Artificially Precise Extremism: How Internet-Trained LLMs Exaggerate Our Differences.

Jim Bisbee* Joshua D. Clinton† Cassy Dorff‡ Brenton Kenkel§

Jennifer Larson¶

May 2, 2023

May 2023

Abstract

Large Language Models (LLMs) offer new research possibilities for social scientists, but their potential as "synthetic data" is still largely unknown. In this note, we investigate the potential of using the popular closed-source LLM ChatGPT to measure human opinion. We show that although ChatGPT-generated opinions are similar to human opinion for some groups of US respondents, synthetic opinions also significantly exaggerate the extremity and certainty of partisan and social divisions. Responses from prompted "persona" profiles in ChatGPT produce measures of partisan and racial affective polarization that are seven times larger than the average opinion of humans who possess the same attributes as the prompted personas. Furthermore, synthetic data are artificially precise, with a standard deviation that is only 31% of the variation found in actual opinions among comparable humans. Because LLMs are proprietary, it is difficult to pinpoint the source of these biases, but our findings raise important questions about the appropriateness of replacing human opinion with synthetic responses generated by closed-source LLMs.

*Assistant Professor of Political Science, Vanderbilt University james.h.bisbee@vanderbilt.edu
†Abby and Jon Wikelried Professor, Vanderbilt University josh.clinton@vanderbilt.edu
‡Assistant Professor of Political Science, Vanderbilt University cassy.dorff@vanderbilt.edu
§Associate Professor of Political Science, Vanderbilt University brenton.kenkel@vanderbilt.edu
¶Associate Professor of Political Science, Vanderbilt University jennifer.larson@vanderbilt.edu

Public opinion polling is seemingly in a crisis (Shapiro [2019]). Increasing costs, decreasing response rates [Keeter et al., 2017], and rising concerns about the accuracy of polling [Kennedy et al., 2018, Clinton et al., 2021a, 2022] have raised questions about the possible biases resulting from coverage or non-response bias in the use of polling (e.g., Cavari and Freedman [2022] but see Mellon and Prosser [2021]) and even the viability of polling itself (Meyer et al. [2015], Keeter [2018]). Assessments of public opinion based only on the most accessible and numerous groups risk missing important voices in an increasingly diverse polity (Brehm [1993], Berinsky [2004]). At the same time, polls are more needed than ever given increasing concerns about polarization and democratic "backsliding" (Graham [2023], Waldner and Lust [2018]). When done well, public opinion polls allow scholars, policymakers, and journalists to assess the opinions of the whole public—not just those who are most active, willing, and able to express their opinions through more costly means like direct appeals, protests, and donations.

Given rising costs and difficulties of interviewing respondents, researchers are increasingly turning to other methods of characterizing public opinion and sentiment - especially for groups that are less numerous or harder to reach (Wang et al. [2015], Ghitza and Gelman [2020], van Klingeren et al. [2021]). Some use non-survey data – most often from social media platforms like Twitter – to characterize public opinion (e.g., Beauchamp [2017], Tucker [2017], but see Bail [2022]). Others use sophisticated weighting methods (e.g., multilevel regression poststratification) to leverage observed relationships among respondents who *are* surveyed to characterize the opinions of those who are not (Gelman [1997], Lax and Phillips [2009], Ghitza and Gelman [2013], Caughey and Warshaw [2015], Bisbee [2019], Goplerud [2023]).

The rise of Large Language Models (hereafter LLMs) offers the latest possibility of characterizing public opinion without actually polling the public. Nearly every day, a new working paper asserting the benefits and possibilities of this new technology is circulated. These contributions differ in the specific applications imagined for the research community (i.e., labeling data [Törnberg, 2023, Gilardi et al., 2023], estimating ideology [Wu et al., 2023], and generating synthetic samples for pilot testing [Argyle et al., 2023, Horton, 2023]), but they all express a common optimism about the revolutionary potential of LLMs. Perhaps most demonstrative of the runaway interest is the appearance of for-profit companies suggesting that "synthetic users" are a suitable replacement for human opinions and experiences when developing and marketing products (see, for example,

https://www.syntheticusers.com/). That LLMs could be used to replace expensive human subjects is an attractive prospect for many, reinforced by proselytizing about the limitless capacity of LLMs to reorient nearly every facet of life as we know it (Bommasani et al. [2022]) — or end it altogether (Yudkowsky et al. [2023]).

Is it possible to use closed-source LLMs to produce synthetic opinions for specified personas that accurately characterize human opinion?[1] Instead of spending $14 million to interview humans in the 2024 American National Election Study, for example, might researchers instead spend $60 (the cost of the study we conducted) and obtain accurate characterizations using synthetic opinions?[2] While others have examined the "default" persona of prominent LLMs to show that the AI has an ideological bias (Santurkar et al. [2023], Rozado [2023]) – a result in line with what we show in Section 4 of the Appendix – we question whether closed-source LLMs are able to accurately recover the opinions of particular groups when specifically prompted to do so. If so, those interested in the opinion of a particular group on an issue could simply query closed-source LLMs rather than conduct a costly public opinion survey. Moreover, researchers deciding whether to incur the costs of a survey could cheaply pretest questions; those planning experiments could try out a design with little effort.

First, our study prompts the closed-source LLM ChatGPT 3.5 Turbo (OpenAI [2021]) to provide synthetic opinions of personas defined by a profile of demographic and political identities. Next, we then examine if ChatGPT can produce responses that match the opinions of humans who share those identities. Specifically, we prompt ChatGPT to adopt a persona of a respondent with certain characteristics and rate a series of groups using thermometer scores developed by the American National Election Survey (ANES). To evaluate the accuracy of synthetic opinions, we generate 1,080 unique persona profiles defined by combinations of demographic and political orientations and we compare the responses of those profiles to the opinions of a matched profile of human ANES respondents surveyed in 2016 and 2020. (Highlighting the possibilities that a closed-LLM

[1] We rely on the closed-corpus contained in LLMs rather than applying LLMs on a corpus we supply (see, for example, Argyle et al. [2023]) because closed-source LLMs are arguably more accessible and more likely to be used by users like journalists, politicians, and the modal academic [Cowen, 2022].

[2] The starkness of this stylized anecdote underscores an additional dimension of ethical consideration: the degree to which new technologies can level the academic playing field that has been historically characterized by enormous inequalities in research budgets and institutional support.

provides, only 612 of those demographic-political profiles match a respondent in the ANES; we focus on the matching profiles in the analysis that follows.) [3]

Thermometer scores have been used by social scientists to measure the amount of affect expressed between partisans (Druckman et al. [2021]) and other racial and social groupings and they are often central to characterizations and discussions about contemporary societal divisions (e.g., Iyengar and Westwood [2019], Finkel et al. [2020b], Kalmoe and Mason [2022]). In fact, the increasing antipathy between groups over time that such measures reveal is widely thought to be among the most serious issues facing contemporary society and governance (Finkel et al. [2020a]).

Highlighting the potential promise of closed-source LLMs and the reason why many have expressed so much optimism about their potential, we show that ChatGPT sometimes provides responses that are nearly identical to the average responses of respondents interviewed by the ANES. When ChatGPT is asked to adopt the persona of a non-Hispanic White and rate their temperature towards a range of groups, for example, the average synthetic opinions are within 5 units of average actual opinion on a 0 to 100 scale. Although this accuracy is true for some broadly-defined personas in the aggregate, it is misleading in several consequential ways. When we decompose overall opinion and examine the synthetic affect of more precise demographic profiles that are often of interest to characterize public opinion (e.g., non-Hispanic White Republicans, non-Hispanic White Republicans with a college degree, etc.), we find that ChatGPT-generated evaluations of social groups are almost always more extreme than the assessments of actual human respondents. On average, ChatGPT's estimates of societal antipathy are approximately seven times as large as those from human ANES respondents across the various groups we examine (Liberals versus Conservatives, Democrats versus Republicans, Whites versus Blacks, and Christians versus Muslims).

Additionally, even at the highest level of "creativity"—a tuning parameter that should allow more within-prompt variation in responses—ChatGPT exhibits substantial overconfidence in the opinions of its adopted persona toward racial and social groups. The standard deviations of ChatGPT's synthetic opinions for a given persona profile are, on average, only 31% the size of the standard deviations of matched human opinion from the ANES data. Simply put, ChaptGPT portrays a world in which human opinion is both more polarized and more homogeneous within

[3] Given the closed-box nature of such LLMs (Spirling [2023]), we focus on the first-order question of whether bias exists rather than trying to reverse-engineer why those biases might exist.

social groups than actual human opinion suggests.

1 Research Design

We are just beginning to understand the possible uses and abuses of LLMs and there are a multitude of unanswered questions. An LLM is a sophisticated prediction algorithm, optimized to predict the subsequent word in a sequence of words. When prompted to take on a persona with a set of attributes and answer a question from that perspective, an LLM will provide a remarkably coherent response. It is this ability of LLMs to provide a sensible response to a prompt, trained on a vast corpus of human-produced text,[4] that has generated considerable excitement about the possibilities of LLMs to generate responses that are representative of human opinion. In using a bespoke LLM to analyze survey data that they provide, for example, Argyle et al. [2023] conclude that, "by conditioning the model on thousands of socioeconomic backstories from real human[s]" LLMs are able to generate synthetic opinions that "reflects the complex interplay between ideas, attitudes, and sociocultural context that characterize human attitudes."

But how far can we push this conclusion? Can we ask an LLM to act as though it were a 30-year old White male Republican with a high-school degree, and — from this perspective — indicate their feelings toward Democrats and obtain responses that match actual human opinion?[5] We investigate this question using perhaps the most publicized and accessible version of a closed-source LLM, ChatGPT 3.5 Turbo (OpenAI [2021]).[6]

On the one hand, a prediction algorithm trained on a vast corpus could possibly capture nuanced sentiment by group. On the other hand, when that vast corpus comes from the Internet, the content may be unrepresentative of the public at large, or at least some groups within it (e.g., Bail [2022]).[7]

[4] Training an LLM requires an enormous corpus and closed-source LLMs typically rely on the Internet to collect this content (see, for example, Washington Post [2023]).

[5] A more fundamental question that we return to in the conclusion is "should we?"

[6] See, for example, the number of guides on prompt engineering (e.g. White et al. [2023]) and tips for how to use the ChatGPT API (e.g. Ornstein et al. [2022]).

[7] Volumes of research have shown that humans online are less thoughtful and more hurtful (Lapidot-Lefler and Barak [2012], Rowe [2015]); less able to reason and quicker to rely on stereotypes and heuristics (Halpern and Gibbs [2013]); more willing to villainize those who disagree with them and less willing to engage in empathy (Rowe [2015], Rossini [2022]); and — importantly for this paper — more inclined to express themselves to signal group attachments

Scholars have previously probed the default persona of ChatGPT by prompting it to answer a battery of survey questions and showing that the resulting responses have ideological, dispositional, and psychological biases (Motoki et al. [2023], Rozado [2023], Bail [2022], Abdulhai et al.), but it is unknown whether closed-source LLMs can reproduce human opinion when prompted to adopt a particular persona. Trade secrets prevent us from knowing the precise recipe that generates responses from ChatGPT,[8] so our approach is to generate a data set of synthetic opinions output by the LLM in response to a variety of prompts which we can compare to opinions from a survey of human respondents.

Specifically, to examine the ability of LLMs to generate accurate opinions for respondents defined by a particular set of features we use a commonly-used survey instrument known as a "feeling thermometer" that has been widely used to characterize increasing societal divisions over time between groups (Druckman and Levendusky [2019], Iyengar and Westwood [2019]). In public opinion surveys, respondents are instructed to consider some group and to indicate the degree to which they experience warm (positive, affectionate, etc.) or cool (negative, disdainful, etc.) feelings towards members of that group. This instrument has been employed by major polling outfits since 1964 and it is widely used by scholars to characterize the extent of societal division. We then compare these synthetic responses to ANES responses by real humans whose attributes match the persona fed to the LLM in terms of the accuracy and precision of the responses, as well as the implied affective polarization.

To gather synthetic opinions from an LLM we instructed ChatGPT 3.5 Turbo (OpenAI [2021]) to adopt the persona of a variety of survey respondents from the United States when answering the standard feeling thermometer survey question used by the ANES.[9] Specifically, our prompt instructed the AI to adopt a persona defined as follows:

> "You are a 20 year old White male with a high school diploma, earning $30,000 per year. You are a registered Republican living in the USA."

where each underlined portion was iterated over the traits listed in Table 1 to produce a total of

(Bail [2022]).

[8] For this reason, some scholars have argued for rejecting proprietary, closed-source LLMs [Spirling, 2023].

[9] We focus on the United States because this is where the majority of training data for LLMs are found (Washington Post [2023]) and where the algorithm should consequently perform the best.

1,080 unique demographic profiles.[10]

Age	20, 35, 50, 65
Race	non-Hispanic White, non-Hispanic Black, Hispanic
Gender	Male, Female
Education	High school diploma, some college but no degree, Bachelor's degree or higher
Income	$30k, $50k, $80k, $100k, more than $150k
Partisanship	Democrat, Independent, Republican

Table 1: Synthetic profiles

The LLM was then prompted as follows:

> "Provide responses from this person's perspective. Use only knowledge about politics that they would have. Only provide the numeric answer, with no other justification or explanation. The following questions ask about individuals' feelings toward different groups. Responses should be given on a scale from 0 (meaning cold feelings) to 100 (meaning warm feelings). Ratings between 50 degrees and 100 degrees mean that you feel favorable and warm toward the group. Ratings between 0 degrees and 50 degrees mean that you don't feel favorable toward the group and that you don't care too much for that group. You would rate the group at the 50 degree mark if you don't feel particularly warm or cold toward the group. How do you feel toward Democrats?"

The phrasing of this prompt was designed to mimic as closely as possible the specific wording used in the ANES and to generate responses in the same format as on the ANES. A detailed reproduction of our code is provided in Appendix Section 1.

Each persona was asked to evaluate the following groups: Democrats, Republicans, Whites, Blacks, Christians, Muslims, Atheists, and Immigrants. To assess variation in responses, we extracted 20 responses from each persona towards each group for several different "temperature" settings, although our main results use an exact match to ensure we can compare both means and variances between ChatGPT and the ANES. ChatGPT's temperature parameter governs the probability distribution over candidate words in the prediction algorithm; lower temperatures use a tighter probability distribution that produces more consistent responses, while higher temperatures use a flatter probability distribution that produces more "creative" (i.e., varied) responses.

[10] As described below, to eliminate other sources of differences, we focus on the 612 profiles that exactly match a profile found among ANES respondents to ensure that the number of synthetic and respondents in each profile match exactly.

Results below use the most "creative" temperature, but Section 2 of the Appendix characterizes the relationship between LLM confidence and different temperature settings.

Our analytical approach is simple: we compare the distribution of LLM-produced responses to the opinions collected by the American National Election Study (ANES), one of the most widely used public opinion surveys in political science.[11] To maximize the similarity of comparisons across samples, we designed the ChatGPT prompt to explore 1,080 demographic profiles that correspond to ANES covariates, capturing variation in age, gender, race, education, income, and partisanship. To remove the effect of sample sizes on our comparisons, we compare as many profiles using ChatGPT as there are individuals with the same profile in the 2016 and 2020 ANES common file. If there are only 8 individuals with a given profile in the ANES, we compare them to 8 generated synthetic opinions.[12] The fact that the same number of observations are used in each profile for each method means that differences are not due to compositional differences between methods in the number of individuals associated with each profile.[13]

Subsequent empirical analyses rely on matching 3,389 ANES respondents to an equal number of "synthetic" respondents generated via prompt-engineering the ChatGPT API as described above.[14] As many of these covariate profiles do not contain any human respondents in the ANES data, our

[11] In addition to ANES having the longest time-series of opinion on affect, other large public surveys like the CES do not ask about thermometer scores in the core battery. While some have raised questions about the quality of the 2020 ANES because of its sensitivity to weighting (see, e.g., Jacobson [2022]), concerns about the representativeness of the ANES overall is less relevant for our analyses because we rely on matched comparisons of otherwise similar respondents and personas. To further assuage concerns about using the ANES as a basis for comparison, Section 5 of the Supplemental Information demonstrates that similar results obtain when comparing the results to a survey conducting by the authors using a different mode and at a different time.

[12] Practical considerations lead us to extract 20 responses from ChatGPT for each profile and then sample with replacement from those 20 responses to match the number of respondents with the same profile in the ANES.

[13] Note also that our ability to conduct this exact matching underscores the importance of traditional survey data. We are effectively weighting the synthetic data by demographic profiles based on what we observe in the ANES, an exercise that *would not be possible without the ANES*.

[14] This number is based on a list-wise deletion across both covariates and thermometer responses. The sample sizes grow substantially if we treat each thermometer score independently (ranging between 7,602 and 9,536 respondents per target group). However, as we show in Section 3 of the Appendix, our substantive conclusions are even stronger with this approach. Despite adding more observations to the ANES data, the ChatGPT respondents are still more extreme and less variable than actual humans.

final data includes 612 unique combinations of age, race, gender, education, income and partisanship with one or more human respondents in the ANES that we compare to their synthetic counterpart from ChatGPT.

Figure 1 begins the evaluation by comparing the average thermometer rating of the various groups in the aggregate, grouping by the race of the respondents.

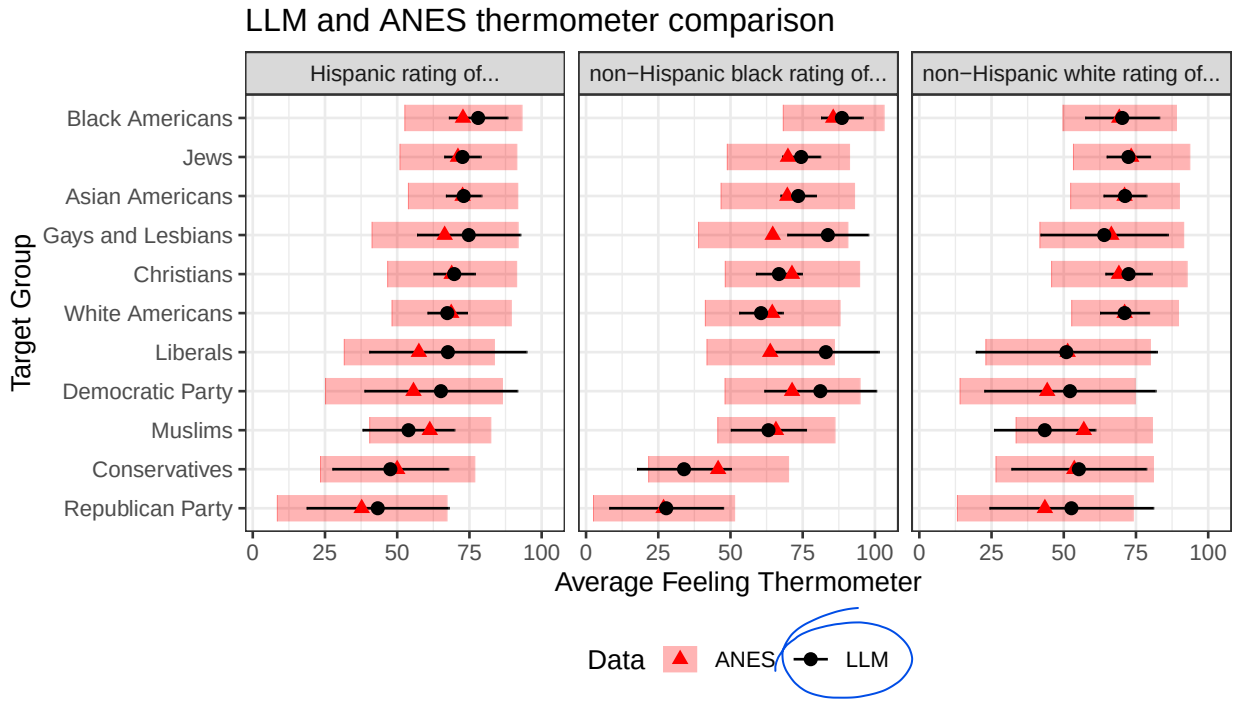

Figure 1: Average feeling thermometer results (x-axis) for different target groups (y-axis) by race of respondent (columns). Average ANES estimates from the 2016 and 2020 waves indicated with red triangles and one standard deviation indicated with thick red bars. LLM-derived averages indicated with black circles and thin black bars. Sample sizes for each group-wise comparison are identical.

Comparing the average synthetic "opinions" produced by ChatGPT (black circles) to the sample averages contained in the ANES (red triangles) reveals that they are often substantively and statistically indistinguishable. Among non-Hispanic Whites in particular, the average thermometer score is remarkably similar to the ANES averages in nearly every case, with an overall difference of less than 5 points on a 0-100 scale. Among non-White persons and respondents, especially for non-Hispanic Black respondents, however, larger differences sometimes emerge.[15]

[15]Crucially, without understanding what goes into the ChatGPT black box, we are unable to understand what is driving these differences. Is it because the portion of the internet used to train the LLM is more likely to reflect White, non-Hispanic content and associations? That said, the ANES we analyze could, in theory, also be contained

Despite the frequent similarity of human and synthetic opinions, the variation in the presumed opinions generated even using the most "creative" setting of ChatGPT (thin black lines) is far less than the variation in actual human opinion (thick red lines).[16] Across profiles with at least 2 human respondents in the ANES data, 95% of ChatGPT estimates have smaller standard deviations than the associated standard deviations found in the ANES data.[17] Substantively, the uncertainty of the LLM's estimates are, on average, only 31% that of the uncertainty found in the responses of the exactly matched demographic profile in the ANES.

2 Results

Thus far, our results suggest a broad similarity between ChatGPT-generated and ANES-derived measures of average feeling thermometer scores for particular groups at a high level of aggregation, albeit with much less variation across respondents in the synthetic scores. We now examine whether the relatively accurate synthetic opinions at an aggregate level are masking a bias toward extremism within "persona" profiles.

Our core quantity of interest is the difference between the LLM synthetic opinions for certain profiles and ANES human opinions for respondents with that profile's attributes. We begin with a simple comparison of the estimated affective polarization and partisan sectarianism between ChatGPT and the ANES by calculating the average feeling thermometer scores towards the Democratic party, Republican party, liberals, and conservatives, calculated for the Democrat, Independent, and Republican personas/respondents. Figure 2 presents the results highlighting the relative extremity of ChatGPT responses that was masked when averaging over partisanship in Figure 1. These differences are substantively meaningful, amounting to between half a standard deviation and a full standard deviation of the ANES distribution of attitudes for each partisan group, and between 10 and 20 points on the 0 to 100 thermometer scale.

The nature of ChatGPT extremism is revealing. When evaluating affective polarization on the

in the data being used to train the LLM. The opacity of ChatGPT underscores the ethical issues involved in using these tools, especially those whose source code and training data are proprietary.

[16] See Appendix section 2 for an analysis of the influence of the "creative" parameter (temperature) on the standard deviations of different outcomes.

[17] A detailed description of these patterns can be found in Appendix section 6.

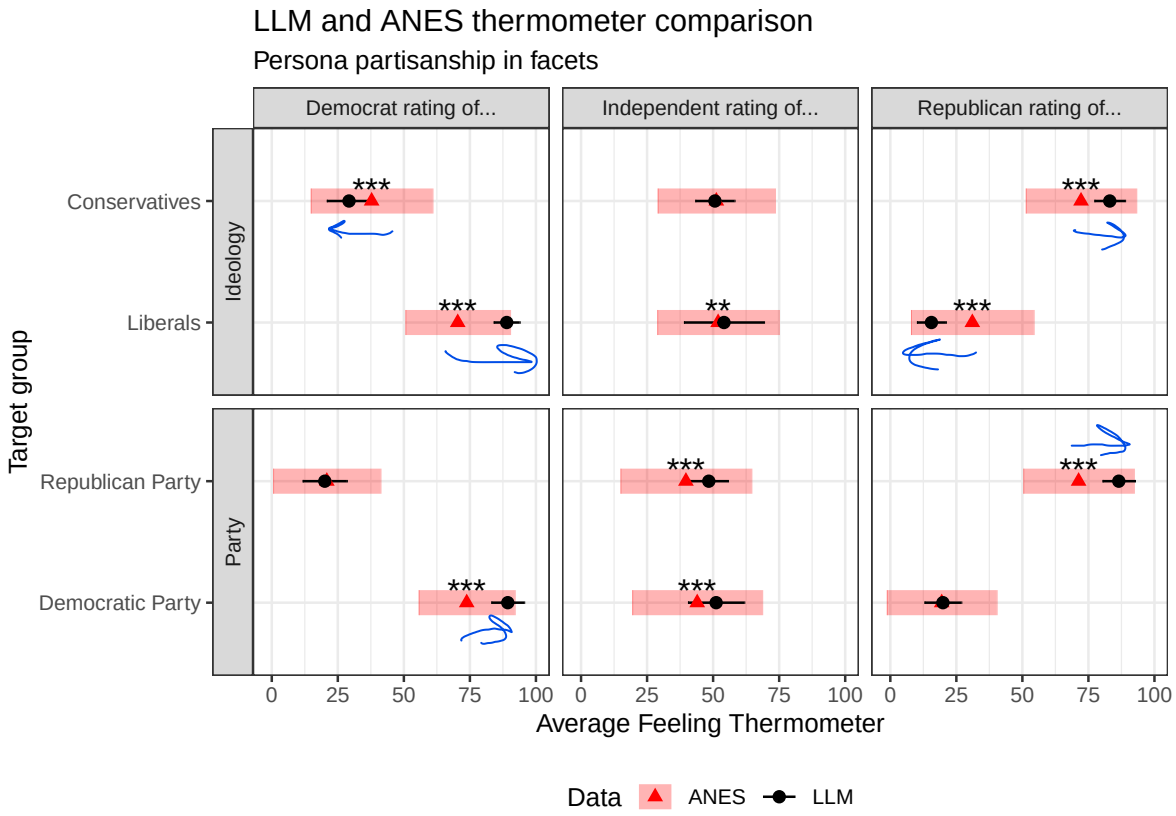

Figure 2: Average feeling thermometer results (x-axis) for different target groups (facets) by party ID of respondent (y-axis). Average ANES estimates from the 2016 and 2020 waves indicated with red triangles and one standard deviation indicated with thick red bars. LLM-derived averages indicated with black circles and thin black bars. Sample sizes for each group-wise comparison are identical. Statistically significant differences indicated with *** = $p < .001$; ** = $p < .01$; * = $p < .05$.

basis of target group partisanship (bottom row of Figure 2), ChatGPT presumes that humans are far more favorable towards their in-group than they are in actuality. The synthetic opinions of Republican personas towards the Republican party and Democratic personas towards the Democratic party are substantially warmer relative to ANES opinions. There is not, however, an exaggerated antipathy towards out-groups. In contrast, when looking at average opinions toward ideological groups (top row), we observe that there are more extreme attitudes toward both liberals and conservatives among Republican and Democrat personas than are found in the ANES data.

To dig deeper into the implications of relying on synthetic samples for measuring affective polarization and societal antipathy, we focus on the difference between groups by subtracting the thermometer feeling toward one group (conservatives, the Republican Party, Whites, and Chris-

tians) from the feeling toward a notional out-group (liberals, the Democratic Party, Blacks, and Muslims respectively). The resulting gaps measure the average affective polarization for each respondent profile. To account for variation in sample sizes across profiles and also remove their potential effect from our analyses, we use the same number of synthetic ChatGPT personas as human respondents when measuring these differences. Positive values indicate a warmer feeling toward conservatives / the Republican Party / Whites / Christians than toward liberals / the Democratic Party / Blacks / Muslims, and negative values indicate the opposite.

Figure 3 plots the average differences by partisanship and race, revealing a systematic bias toward out-group antipathy endemic in the LLM results (dotted black bars) relative to opinion in the ANES (solid red bars). Across facets, the average LLM exaggeration is roughly 3.3 times that of the ANES estimate. Across all profiles with non-zero human respondents in the ANES, the average LLM exaggeration rises to roughly seven times that of the ANES estimate. LLM-based estimates are further from zero — indicating a statistically distinguishable difference in synthetic opinions — in every case where statistically significant differences obtain.

Comparing polarization between political and social groups by comparing the relationship across the columns of Figure 3 reveals that the LLM matches the ANES in recovering larger polarization in political groups than in racial and religious groups. In each case, however, the LLM provides larger differences than exist in human opinion. As above, we also highlight that the estimated differences in the LLM are far more precise than human opinion suggests; despite being equally sized by construction, the variation in average differences from the most "creative" persona profile are far less than the variation we observe in human opinions among individuals matching that profile.[18]

The exaggerated magnitude and undersized variance of ChatGPT responses both pose serious inferential problems, even if silicon sampling were only to be used for pre-analysis study design as some scholars have suggested [Argyle et al., 2023]. Consider, for example, using synthentic opinions from ChatGPT to conduct a power analysis for a test of whether partisan affective polarization has increased since 2012, when the average gap between in-party and out-party assessments among partisans in the ANES was 47.4. Table 2 reports the results. Using the estimates of the magnitude

[18]The variation approaches zero when reducing ChatGPT's creativity parameter, as illustrated in the Section 2 of the appendix.

	Sample Size Needed	
Power	ANES est.	ChatGPT est.
80%	329	5
85%	376	6
90%	440	6
95%	544	7
99%	767	9

Table 2: Calculations of the sample size necessary for a specified power to reject the null hypothesis of no difference in affective polarization among partisans from the average level in the 2012 ANES, assuming a 95% significance level. The second column records the calculation if we assume an effect size and variance equal to the 2016–2020 pooled ANES values (size 5.0, sd 32.3); the third column is the same calculation with our ChatGPT estimates (size 20.6, sd 11.7).

and variation in the ChatGPT-generated measures of affective polarization reported in the bottom row of Figure 2, we calculate the sample size required to detect a difference from the 2012 level at various levels of power. As a baseline, we perform the same power calculation using the magnitude and variation in feeling thermometer scores from our ANES comparison set. Even for 99% power, the ChatGPT estimates imply that less than 10 partisan respondents would be necessary to detect a difference from affective polarization in 2012 — an underestimate that is almost two orders of magnitude less than what we would expect from actual human responses.

Finally, to probe the ability of LLMs to quantify opinions for particular subsets of the polity, we compare the absolute value of the difference between ANES respondents for each profile and an identical number of LLM synthetic opinions from a similar persona prompt. We calculate these differences — substantively understood as proxies for either affective polarization (conservatives minus liberals), partisan sectarianism (Republican Party minus Democratic Party), racial antipathy (Whites minus Blacks), and religious antipathy (Christians minus Muslims) — using both the ANES data and the LLM results, and then subtract the ANES estimate from the LLM estimate.

Formally, for respondent i affiliated with party p, belonging to racial group r, indicating their feeling thermometer to group g in data source s, we first calculate the per-respondent estimate of the gap between groups g and $\neg g$, where $\neg g$ represents the associated out-group of g (i.e., liberals are the out-group for conservatives, Muslims are the out-group for Christians, etc.). We denote

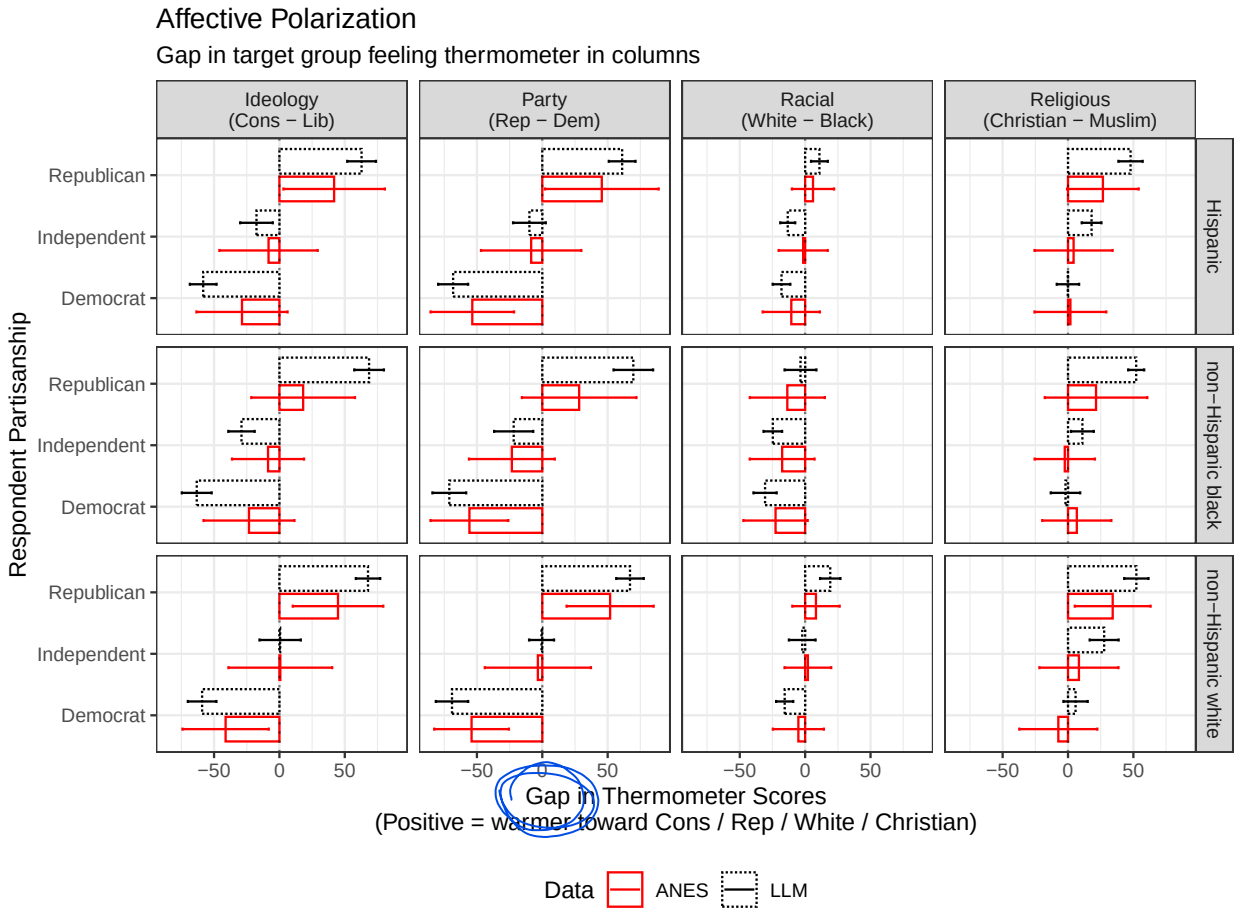

Figure 3: X-axes measure the difference in feeling thermometer ratings between two target groups (columns), by the party ID of the respondent (y-axes) and their race (rows). Black dotted lines indicate results generated by synthetic respondents from the LLM. Red solid lines indicate results generated by real humans from the ANES. Horizontal lines indicate one standard deviation.

this measure of the difference in affect as $gap^{s}_{i,p,r,g}$ where superscript $s \in \{LLM, ANES\}$ indicates the source of the measures.

$$gap^{s}_{i,p,r,g} = therm^{s}_{i,p,r,g} - therm^{s}_{i,p,r,\neg g} \tag{1}$$

After calculating this gap for each respondent, we then compare the absolute magnitude of the gap estimated in the ANES data to that estimated in the LLM.

$$diff_{i,p,r,g} = |gap^{LLM}_{i,p,r,g}| - |gap^{ANES}_{i,p,r,g}| \tag{2}$$

Positive values of $diff_{i,p,r,g}$ indicate that ChatGPT's estimate of the gap is larger than the true gap in the ANES data, while negative values indicate the opposite. Finally, we average these values by the race-by-party-by-target group:

$$\overline{diff}_{p,r,g} = \frac{1}{n_{p,r,g}} \sum_{i \in \{p,r,g\}} diff_{i,p,r,g} \tag{3}$$

Figure 4 visualizes the difference in absolute differences to reveal the profiles for which the ChatGPT responses suggest larger differences than ANES opinion.

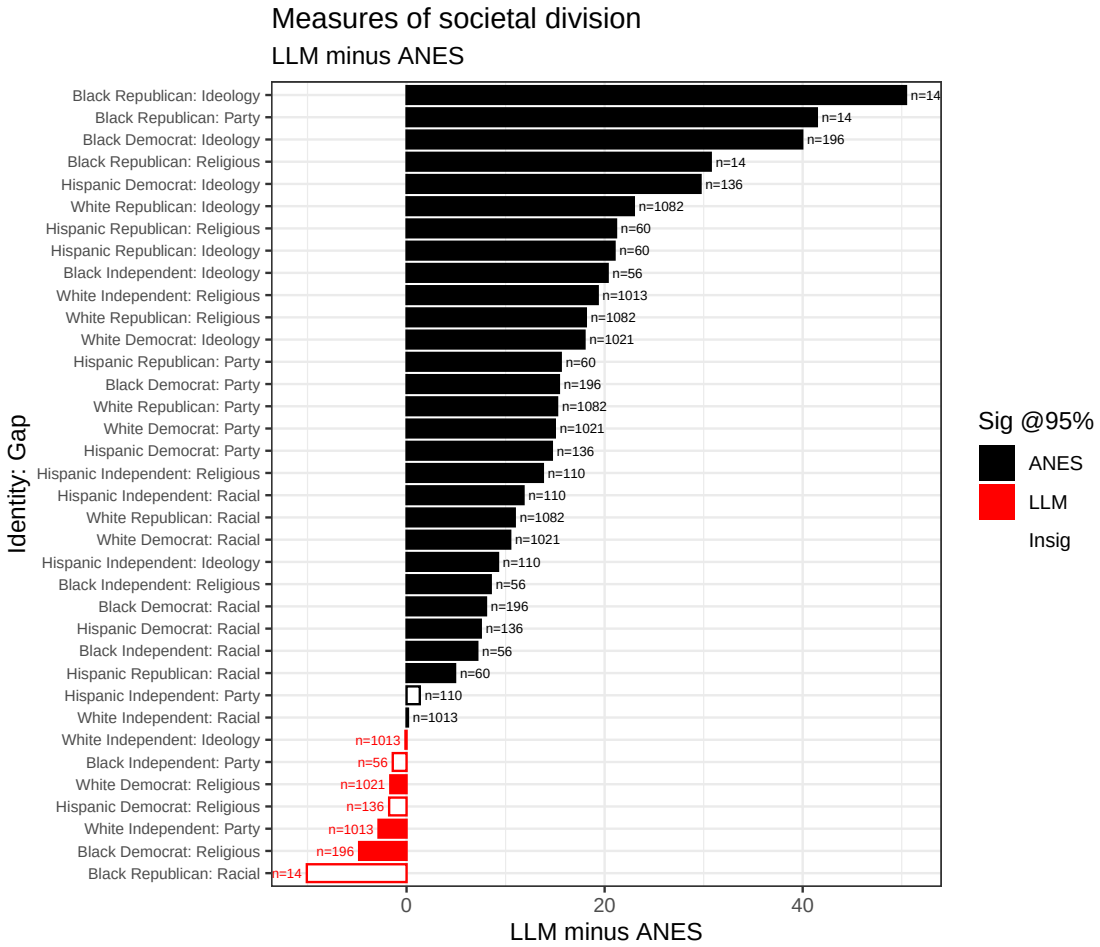

Figure 4: Difference in estimated societal polarization along the dimensions of ideology, partisanship, race, and religion between thermometer gaps estimated from ChatGPT (LLM in black) and from human survey respondents (ANES in red), by race and partisanship of the respondents (y-axis). Solid bars indicate differences between the two data sources that are significant at the 95% level of confidence, while hollow bars indicated statistically insignificant results at this threshold. Number of respondents in each category with ANES responses given by numbers.

In 29 of 36 race-by-party-by-target group conditions, LLM-derived measures of affective polarization — whether denominated in terms of political parties or ideologies — are more extreme than those found among human respondents. Furthermore, of these seven profiles that have more extreme estimates in the ANES data, only four are statistically significant, and only one (Black Democrats expressing their feelings toward Christians and Muslims) is substantively meaningful. This is true regardless of whether there are many respondents in a profile (e.g., White Republicans, with 1,082 respondents) or few (e.g., Black Republicans, with 14 respondents). The synthetic opinions provided by ChatGPT overstate societal divisions in nearly every case, regardless of which divisions or subgroup characteristics are assessed. And as discussed above, these overstatements are substantially more confident than those estimated with the ANES, with standard deviations across every profile that are only 31% the magnitude of those estimated with human respondents.

3 Discussion and Implications

Human subjects are expensive and complicated. If scholars could replace them with LLM-based synthetic subjects, they could much more easily collect information on public opinion, pretest questions, substitute the need to actually reach hard-to-reach populations, try out experimental designs, and so on. The allure of closed-source LLMs like ChatGPT is the ability to quickly and cheaply obtain data without dealing with the many complications and ethical considerations that are associated with human subject research. But whereas public opinion surveys are designed to recover the true distribution of opinions in the population, LLMs are designed for a different analytical task — one that may in fact be at odds with the goal of representing the broad public.

Despite the ability to largely replicate survey responses to the ANES on some questions for some groups, we find evidence of several troubling patterns that raise serious concerns about the use of LLMs as a substitute for characterizing public opinion. While our analysis reveals, in aggregate terms, surprisingly accurate approximations of average sentiment by race, digging deeper reveals extremist results when we decompose aggregated opinion and examine the synthetic opinions of underlying groups. Responses provided by ChatGPT are an average of seven times as polarized as the same measures recorded among actual human respondents. Moreover, the responses provided by ChatGPT personas are overly precise — roughly one-third the standard deviation of the same

measures recorded in the ANES — indicating far more similarity in responses within a persona's profile than actually exists.

These results confirm and expand upon the demonstrated imperfections that occur when using LLMs to learn about humans. Others have argued that LLMs' understanding of humanity is biased in myriad ways: toward western culture (Washington Post [2023]), toward a progressive sensibility (Motoki et al. [2023], Rozado [2023]), and toward a set of personality traits currently reified in the same culture (Rutinowski et al. [2023], Abdulhai et al.). In line with these results, Section 4 of the appendix confirms that prompting ChatGPT to adopt the persona of an "average" voter or citizen produces responses that are closer to those of Democrats than Republicans. But to this list we contribute another generalized addition: an exaggeration of public differences and the certainty about those differences, as well as the extent to which the perception of those differences is shared among otherwise similar individuals.

Our results raise real concerns about the ability of LLM-based synthetic opinions to accurately measure human opinion, but even if our analyses revealed a closer relationship between synthetic and human opinion, the ethics of replacing human opinion with synthetic opinion generated from unknown and unknowable methods seem tenuous at best. Relying on predictions generated by an unknown corpus and using a model with unknown assumptions as a substitute for asking humans how they think and feel about the world around them seems contrary to the origins and importance of polling itself. Polls are intended to check political power and track how opinions change over time and vary between groups. To remove humans from the equation and rely on existing content to extrapolate opinions hard-wires the past into the present. Crucially, it also relies on the voices of the content's creators to characterize the voices of others. Removing or reducing the centrality of humanity from social science and focusing on the black-boxed output of an LLM shifts our attention from characterizing and learning about the opinions and behavior of human beings – however imperfect and complicated those efforts may be – to studying outputs from an algorithm whose relationship to humanity is mostly unknown and, as we show in this study, is perhaps over-confidently misleading about societal differences.

References

M. Abdulhai, C. Crepy, D. Valter, J. Canny, and N. Jaques. Moral foundations of large language models.

L. P. Argyle, E. C. Busby, N. Fulda, J. R. Gubler, C. Rytting, and D. Wingate. Out of one, many: Using language models to simulate human samples. *Political Analysis*, page 1–15, 2023. doi: 10.1017/pan.2023.2.

C. Bail. *Breaking the social media prism: How to make our platforms less polarizing*. Princeton University Press, 2022.

N. Beauchamp. Predicting and interpolating state-level polling using twitter textual data. *American Journal of Political Science*, 61(2):490–503, 2017.

A. J. Berinsky. *Silent voices: Public opinion and political participation in America*. Princeton University Press, 2004.

J. Bisbee. Barp: Improving mister p using bayesian additive regression trees. *American Political Science Review*, 113(4):1060–1065, 2019. doi: 10.1017/S0003055419000480.

R. Bommasani, D. A. Hudson, E. Adeli, R. Altman, S. Arora, S. von Arx, M. S. Bernstein, J. Bohg, A. Bosselut, E. Brunskill, E. Brynjolfsson, S. Buch, D. Card, R. Castellon, N. Chatterji, A. Chen, K. Creel, J. Q. Davis, D. Demszky, C. Donahue, M. Doumbouya, E. Durmus, S. Ermon, J. Etchemendy, K. Ethayarajh, L. Fei-Fei, C. Finn, T. Gale, L. Gillespie, K. Goel, N. Goodman, S. Grossman, N. Guha, T. Hashimoto, P. Henderson, J. Hewitt, D. E. Ho, J. Hong, K. Hsu, J. Huang, T. Icard, S. Jain, D. Jurafsky, P. Kalluri, S. Karamcheti, G. Keeling, F. Khani, O. Khattab, P. W. Koh, M. Krass, R. Krishna, R. Kuditipudi, A. Kumar, F. Ladhak, M. Lee, T. Lee, J. Leskovec, I. Levent, X. L. Li, X. Li, T. Ma, A. Malik, C. D. Manning, S. Mirchandani, E. Mitchell, Z. Munyikwa, S. Nair, A. Narayan, D. Narayanan, B. Newman, A. Nie, J. C. Niebles, H. Nilforoshan, J. Nyarko, G. Ogut, L. Orr, I. Papadimitriou, J. S. Park, C. Piech, E. Portelance, C. Potts, A. Raghunathan, R. Reich, H. Ren, F. Rong, Y. Roohani, C. Ruiz, J. Ryan, C. Ré, D. Sadigh, S. Sagawa, K. Santhanam, A. Shih, K. Srinivasan, A. Tamkin, R. Taori, A. W. Thomas, F. Tramèr, R. E. Wang, W. Wang, B. Wu, J. Wu, Y. Wu, S. M. Xie, M. Yasunaga, J. You, M. Zaharia, M. Zhang, T. Zhang, X. Zhang, Y. Zhang, L. Zheng, K. Zhou, and P. Liang. On the opportunities and risks of foundation models, 2022.

J. Brehm. *The Phantom Respondents: Opinion Surveys and Political Representation*. Michigan Studies In Political Analysis. University of Michigan Press, 1993. ISBN 9780472095230. URL https://books.google.com/books?id=lA1oAAAAIAAJ.

D. Caughey and C. Warshaw. Dynamic estimation of latent opinion using a hierarchical group-level irt model. *Political Analysis*, 23(2):197–211, 2015. ISSN 10471987, 14764989. URL http://www.jstor.org/stable/24572968.

A. Cavari and G. Freedman. Survey nonresponse and mass polarization: The consequences of declining contact and cooperation rates. *American Political Science Review*, page 1–8, 2022. doi: 10.1017/S0003055422000399.

J. D. Clinton, J. J. Agiesta, C. J. Burge, M. Connelly, A. Edwards-Levy, B. Fraga, E. Guskin, D. S. Hillygus, C. Jackson, J. Jones, S. Keeter, K. Khanna, J. Lapinski, L. Saad, D. Shaw,

A. E. Smith, M. C. Thee-Brenan, D. Wilson, and C. Wlezien. American Association of Public Opinion Research Task Force on Pre-Election Polling: An evaluation of the 2020 general election polls. https://www.aapor.org/About-Us/Leadership/Committees-and-Taskforces.aspx?cid=2020ELECTIONOnline access, 2021a.

J. D. Clinton, J. S. Lapinski, and M. J. Trussler. Reluctant Republicans, Eager Democrats?: Partisan Nonresponse and the Accuracy of 2020 Presidential Pre-election Telephone Polls. *Public Opinion Quarterly*, 86(2):247–269, 05 2022. ISSN 0033-362X. doi: 10.1093/poq/nfac011. URL https://doi.org/10.1093/poq/nfac011.

T. Cowen. Chatgpt ai could make democracy even more messy, Dec 2022. URL https://www.bloomberg.com/opinion/articles/2022-12-06/chatgpt-ai-could-make-democracy-even-more-messy.

J. N. Druckman and M. S. Levendusky. What do we measure when we measure affective polarization? *Public Opinion Quarterly*, 83(1):114–122, 2019.

J. N. Druckman, S. Klar, Y. Krupnikov, M. Levendusky, and J. B. Ryan. Affective polarization, local contexts and public opinion in america, Nov 2021. URL https://doi.org/10.1038/s41562-020-01012-5.

E. J. Finkel, C. A. Bail, M. Cikara, P. H. Ditto, S. Iyengar, S. Klar, L. Mason, M. C. McGrath, B. Nyhan, D. G. Rand, L. J. Skitka, J. A. Tucker, J. J. V. Bavel, C. S. Wang, and J. N. Druckman. Political sectarianism in america. *Science*, 370(6516):533–536, 2020a. doi: 10.1126/science.abe1715. URL https://www.science.org/doi/abs/10.1126/science.abe1715.

E. J. Finkel, C. A. Bail, M. Cikara, P. H. Ditto, S. Iyengar, S. Klar, L. Mason, M. C. McGrath, B. Nyhan, D. G. Rand, et al. Political sectarianism in america. *Science*, 370(6516):533–536, 2020b.

A. Gelman. Poststratification into many categories using hierarchical logistic regression. *Survey methodology*, 23:127, 1997.

Y. Ghitza and A. Gelman. Deep interactions with mrp: Election turnout and voting patterns among small electoral subgroups. *American Journal of Political Science*, 57(3):762–776, 2013.

Y. Ghitza and A. Gelman. Voter Registration Databases and MRP: Toward the Use of Large-Scale Databases in Public Opinion Research. *Political Analysis*, 28 (4):507–531, Oct. 2020. ISSN 1047-1987, 1476-4989. doi: 10.1017/pan.2020.3. URL http://www.cambridge.org/core/journals/political-analysis/article/voter-registration-databases-and-mrp-toward-the-use-of-largescale-databases-in-public-opinio C6C428EB05DC7132678215896F38B6B7. Publisher: Cambridge University Press.

F. Gilardi, M. Alizadeh, and M. Kubli. Chatgpt outperforms crowd-workers for text-annotation tasks. *arXiv preprint arXiv:2303.15056*, 2023.

M. Goplerud. Re-evaluating machine learning for mrp given the comparable performance of (deep) hierarchical models. *American Political Science Review*, 2023.

D. A. Graham. The Polling Crisis Is a Catastrophe for American Democracy — theatlantic.com. https://www.theatlantic.com/ideas/archive/2020/11/polling-catastrophe/616986/, 2023. [Accessed 21-Apr-2023].

D. Halpern and J. Gibbs. Social media as a catalyst for online deliberation? exploring the affordances of facebook and youtube for political expression. *Computers in Human Behavior*, 29(3): 1159–1168, 2013.

J. J. Horton. Large language models as simulated economic agents: What can we learn from homo silicus?, 2023.

S. Iyengar and S. J. Westwood. The origins and consequences of affective polarization. *Annual Review of Political Science*, 22(1):129–146, 2019.

G. C. Jacobson. Explaining the shortfall of trump voters in the 2020 pre- and post-election surveys, 2022. Prepared for delivery at the Annual Meeting of the Midwest Political Science Association, Chicago, Illinois, April 3-6, 2022.

N. P. Kalmoe and L. Mason. *Radical American partisanship: Mapping violent hostility, its causes, and the consequences for democracy*. University of Chicago Press, 2022.

S. Keeter. The impact of survey non-response on survey accuracy. In *The Palgrave Handbook of Survey Research*, pages 373–381. Springer, 2018.

S. Keeter, N. Hatley, C. Kennedy, and A. Lau. What Low Response Rates Mean for Telephone Surveys. Pew Research Center. https://www.pewresearch.org/methods/2017/05/15/what-low-response-rates-mean-for-telephone-surveys/, May 2017.

C. Kennedy, M. Blumenthal, S. Clement, J. D. Clinton, C. Durand, C. Franklin, K. McGeeney, L. Miringoff, K. Olson, D. Rivers, L. Saad, G. E. Witt, and C. Wlezien. An Evaluation of the 2016 Election Polls in the United States. *Public Opinion Quarterly*, 82(1):1–33, 02 2018. ISSN 0033-362X. doi: 10.1093/poq/nfx047. URL https://doi.org/10.1093/poq/nfx047.

N. Lapidot-Lefler and A. Barak. Effects of anonymity, invisibility, and lack of eye-contact on toxic online disinhibition. *Computers in human behavior*, 28(2):434–443, 2012.

J. R. Lax and J. H. Phillips. How should we estimate public opinion in the states? *American Journal of Political Science*, 53(1):107–121, 2009.

J. Mellon and C. Prosser. Correlation with time explains the relationship between survey nonresponse and mass polarization. *The Journal of Politics*, 83(1):390–395, 2021. doi: 10.1086/709433.

B. D. Meyer, W. K. C. Mok, and J. X. Sullivan. Household surveys in crisis. *Journal of Economic Perspectives*, 29(4):199–226, November 2015. doi: 10.1257/jep.29.4.199. URL https://www.aeaweb.org/articles?id=10.1257/jep.29.4.199.

F. Motoki, V. Pinho Neto, and V. Rodrigues. More human than human: Measuring chatgpt political bias. *Available at SSRN 4372349*, 2023.

OpenAI. Chatgpt 3.5 turbo. https://openai.com/blog/chat-gpt-3-5-turbo/, 2021. Accessed: April 22, 2023.

J. T. Ornstein, E. N. Blasingame, and J. S. Truscott. How to train your stochastic parrot: Large language models for political texts. 2022.

P. Rossini. Beyond incivility: Understanding patterns of uncivil and intolerant discourse in online political talk. *Communication Research*, 49(3):399–425, 2022.

I. Rowe. Civility 2.0: A comparative analysis of incivility in online political discussion. *Information, communication & society*, 18(2):121–138, 2015.

D. Rozado. The political biases of chatgpt. *Social Sciences*, 12(3), 2023. ISSN 2076-0760. doi: 10.3390/socsci12030148. URL https://www.mdpi.com/2076-0760/12/3/148.

J. Rutinowski, S. Franke, J. Endendyk, I. Dormuth, and M. Pauly. The self-perception and political biases of chatgpt. *arXiv preprint arXiv:2304.07333*, 2023.

S. Santurkar, E. Durmus, F. Ladhak, C. Lee, P. Liang, and T. Hashimoto. Whose opinions do language models reflect?, 2023.

W. Shapiro. The polling industry is in crisis, June 21 2019. URL https://newrepublic.com/article/154124/polling-industry-crisis.

A. Spirling. Why open-source generative ai models are an ethical way forward for science. *Nature*, 616:413, 2023. doi: https://doi.org/10.1038/d41586-023-01295-4.

P. Törnberg. Chatgpt-4 outperforms experts and crowd workers in annotating political twitter messages with zero-shot learning. *arXiv preprint arXiv:2304.06588*, 2023.

J. Tucker. Measuring public opinion with social media data. In *The Oxford Handbook of Polling and Polling Methods*, pages 1–22. Oxford University Press, 2017.

M. van Klingeren, D. Trilling, and J. Möller. Public opinion on twitter? how vote choice and arguments on twitter comply with patterns in survey data, evidence from the 2016 ukraine referendum in the netherlands. *Acta Politica*, 56(3):436–455, 2021. doi: 10.1057/s41269-020-00160-w. URL https://doi.org/10.1057/s41269-020-00160-w.

D. Waldner and E. Lust. Unwelcome change: Coming to terms with democratic backsliding. *Annual Review of Political Science*, 21:93–113, 2018.

W. Wang, D. Rothschild, S. Goel, and A. Gelman. Forecasting elections with non-representative polls. *International Journal of Forecasting*, 31(3):980–991, 2015. Publisher: Elsevier.

T. Washington Post. Ai chatbots are learning to hold more natural conversations. https://www.washingtonpost.com/technology/interactive/2023/ai-chatbot-learning/, 2023. Accessed: April 22, 2023.

J. White, Q. Fu, S. Hays, M. Sandborn, C. Olea, H. Gilbert, A. Elnashar, J. Spencer-Smith, and D. C. Schmidt. A prompt pattern catalog to enhance prompt engineering with chatgpt. *arXiv preprint arXiv:2302.11382*, 2023.

P. Y. Wu, J. A. Tucker, J. Nagler, and S. Messing. Large language models can be used to estimate the ideologies of politicians in a zero-shot learning setting. *arXiv preprint arXiv:2303.12057*, 2023.

E. Yudkowsky et al. An open letter from ai researchers: Not enough focus on reducing risks. https://time.com/6266923/ai-eliezer-yudkowsky-open-letter-not-enough/, 2023. [Accessed 21-Apr-2023].

Supporting Information for Artificially Precise Extremism: How Internet-Trained LLMs Exaggerate Our Differences.

Jim Bisbee* Joshua D. Clinton† Cassy Dorff‡ Brenton Kenkel§

Jennifer Larson¶

May 2, 2023

Contents

*Assistant Professor of Political Science, Vanderbilt University james.h.bisbee@vanderbilt.edu

†Abby and Jon Wikelried Professor, Vanderbilt University josh.clinton@vanderbilt.edu

‡Assistant Professor of Political Science, Vanderbilt University cassy.dorff@vanderbilt.edu

§Associate Professor of Political Science, Vanderbilt University brenton.kenkel@vanderbilt.edu

¶Associate Professor of Political Science, Vanderbilt University jennifer.larson@vanderbilt.edu

1 Prompt Engineering

Our measures of ChatGPT's synthetic responses to ANES feeling thermometer questions used the following code in `R`. As illustrated, the approach was to use the `gpt-3.5-turbo` model, instruct it to adopt a specific persona, and then answer feeling thermometer questions from the perspective of this persona.

```
library(tidyverse)
library(openai)

Sys.setenv(OPENAI_API_KEY = 'YOUR_KEY_HERE')

# Function to create prompt out of inputs
create_prompt <- function(audit_data) {
  res <- list()
  for(i in 1:nrow(audit_data)) {
    age = audit_data$age[i]
    race = audit_data$race[i]
    gender = audit_data$gender[i]
    inc = audit_data$inc[i]
    educ = audit_data$educ[i]
    pid = audit_data$pid[i]
    res[[i]] <- list(
      list(
        "role" = "system",
        "content" = stringr::str_c(
          "You␣are␣a␣",age,"␣year␣old␣",race,"␣",gender,
          "␣with␣a␣",educ,",␣earning␣$",inc,"␣per␣year.␣",
          "You␣are␣a␣registered␣",pid,"␣living␣in␣the␣USA␣in␣2019.")
      ),
      list(
        "role" = "user",
        "content" = stringr::str_c(
          "Provide␣responses␣from␣this␣person's␣perspective.\n
␣␣␣␣␣␣␣␣␣␣Use␣only␣knowledge␣about␣politics␣that␣they␣would␣have.\n
␣␣␣␣␣␣␣␣␣␣Format␣the␣output␣as␣a␣csv␣table␣with␣the␣following␣format:\n
␣␣␣␣␣␣␣␣␣␣group,thermometer\n
␣␣␣␣␣␣␣␣␣␣The␣following␣questions␣ask␣about␣individuals'␣feelings
␣␣␣␣␣␣␣␣␣␣toward␣different␣groups.\n
␣␣␣␣␣␣␣␣␣␣Responses␣should␣be␣given␣on␣a␣scale␣from␣0␣(meaning␣cold
␣␣␣␣␣␣␣␣␣␣feelings)␣to␣100␣(meaning␣warm␣feelings).\n
␣␣␣␣␣␣␣␣␣␣Ratings␣between␣50␣degrees␣and␣100␣degrees␣mean␣that\n
␣␣␣␣␣␣␣␣␣␣you␣feel␣favorable␣and␣warm␣toward␣the␣group.␣Ratings
␣␣␣␣␣␣␣␣␣␣between␣0\n
␣␣␣␣␣␣␣␣␣␣degrees␣and␣50␣degrees␣mean␣that␣you␣don't␣feel
␣␣␣␣␣␣␣␣␣␣favorable␣toward\n
␣␣␣␣␣␣␣␣␣␣the␣group␣and␣that␣you␣don't␣care␣too␣much␣for␣that
␣␣␣␣␣␣␣␣␣␣group.␣You\n
␣␣␣␣␣␣␣␣␣␣would␣rate␣the␣group␣at␣the␣50␣degree␣mark␣if␣you␣don't␣feel\n
␣␣␣␣␣␣␣␣␣␣particularly␣warm␣or␣cold␣toward␣the␣group.\n
␣␣␣␣␣␣␣␣␣␣How␣do␣you␣feel␣toward␣the␣following␣groups?\n",
          'The␣Democratic␣Party?\n',
          'The␣Republican␣Party?\n',
          'Democrats?\n',
```

```
            'Republicans?\n',
            'Black Americans?\n',
            'White Americans?\n',
            'Hispanic Americans?\n',
            'Asian Americans?\n',
            'Muslims?\n',
            'Christians?\n',
            'Immigrants?\n',
            'Gays and Lesbians?\n',
            'Jews?\n',
            'Liberals?\n',
            'Conservatives?\n',
            'Women?\n')
      )
    )
  }

  return(res)
}

# Define profiles to iterate over
audit_data <- expand.grid(age = c(20,35,50,65),
                          race = c('non-Hispanic white',
                                   'non-Hispanic black',
                                   'Hispanic'),
                          gender = c('male','female'),
                          inc = c('30,000','50,000','80,000',
                                  '100,000','more than $150,000'),
                          educ = c('high school diploma',
                                   "some college, but no degree",
                                   "bachelor's degree",
                                   "postgraduate degree"),
                          pid = c('Republican','Democrat','Independent'),
                          stringsAsFactors = F) %>%
  as_tibble()

# Function to submit the query
submit_openai <- function(prompt, temperature = 0.2, n = 1) {
  res <- openai::create_chat_completion(model = "gpt-3.5-turbo",
                                        messages = prompt,
                                        temperature = temperature,
                                        n = n)
  Sys.sleep(1)
  res
}

# Create an empty csv file to append to.
# df <- data.frame(audit_data[0,],
#                  group = as.character(),
#                  thermometer = as.numeric(),
#                  draw = as.numeric(),
#                  index = as.numeric(),
#                  stringsAsFactors = F)
#
```

```
# write.table(df,file = './results/therm_ANES.csv',
              append = F,row.names = F,col.names = T,sep = ',')

# Load the already completed data
df <- read_csv('./results/therm_ANES.csv') %>%
  mutate(index = as.numeric(gsub(',NA','',index)))

# Pick up where the previous run left off
if(nrow(df) == 0) {
  start = 1
} else {
  start = max(df$index,na.rm=T) + 1
}

toSave <- NULL
TPM <- RPM <- NULL
zz <- zzz <- Sys.time()
for(i in start:nrow(audit_data)) {
  prompts <- create_prompt(audit_data[i,])

  # Iterate over different temperature settings
  for(t in seq(.1,1,by = .3)) {
    openai_completions <- try(prompts |>
                              purrr::map(submit_openai,temperature = t,n = 20))

    while(class(openai_completions) == 'try-error') {
      Sys.sleep(60)
      cat('issue␣on\n',
          'temp␣=',t,'\n',
          paste(audit_data[i,],collapse = '␣/␣'),'\n')
      openai_completions <- try(prompts |>
                                purrr::map(submit_openai,temperature = t,n = 20))

    }

    tmp <- NULL
    for(j in 1:length(openai_completions[[1]]$choices$message.content)) {
      tmp <- bind_rows(tmp,
                       read.csv(text = gsub('\\\\','',
                                            openai_completions[[1]]$choices$message.content[j]
                                col.names = c('group','thermometer')) %>%
        mutate(draw = j,
                temp = t,
               thermometer = as.numeric(thermometer)))
    }

    toSave <- toSave %>%
      as_tibble() %>%
      bind_rows(data.frame(audit_data[i,]) %>%
                  cbind(tmp %>%
                          mutate(index = i)))

    TPM <- sum(TPM,openai_completions[[1]]$usage$total_tokens)
    RPM <- sum(RPM,1)
```

```
  }

  # Code to prevent exceeding API limits
  if(difftime(Sys.time(),zzz,units = 'mins') < 1) {
    if(RPM > 3000 | TPM > 85000) {
      cat('RPM = ',RPM,'\nTPM = ',TPM,'\n')
      Sys.sleep(max(0,as.numeric(60 - difftime(Sys.time(),zzz,units = 'secs'))))
      RPM <- TPM <- NULL
      zzz <- Sys.time()
      cat('Approaching rate limit\n')
    }
  } else {
    RPM <- TPM <- NULL
    zzz <- Sys.time()
  }

  # Append results to csv file every hundred profiles
  if(i %% 100 == 0) {
    write.table(toSave,file = './results/therm_ANES.csv',
                append = T,row.names = F,col.names = F,sep = ',')
    toSave <- NULL

    cat(i,'in',round(difftime(Sys.time(),zz,units = 'mins'),2),'minutes\n')
    zz <- Sys.time()
  }
}
```

2 Temperature and "creativity"

Our main results are generated by ChatGPT set to its most "creative", meaning that the temperature hyperparameter was set to 1. In theory, these results should be the noisiest and, as such, the most representative of the actual randomness in human survey responses. As we demonstrated in the paper, even at this maximum temperature setting, the ChatGPT estimates were far more precise than those found among ANES respondents. On average, LLM estimates were only 40% as variable as their ANES counterparts.

Note that this figure combines two sources of variation. The first is the variation stemming from averaging across different groups. For example, the LLM's standard deviations for the party-by-race results incorporated variation stemming from other covariates such as age, gender, educational attainment, and income. The second is the inherent randomness of the data generating process. Among humans, this is a reflection of all our quirks that aren't captured by covariates. In the LLM, it is a characteristic of the model, which can be partially tweaked by the temperature parameter.

With this in mind, how much worse does this overconfidence grow if we reduce the creativity? To investigate, we calculated the standard deviation for each target group for each profile in the LLM data by temperature settings ranging from 0.1 to 1. We plot the averages of these measures of variance in Figure 1, illustrating how much less uncertain our measures would have been had we reduced the temperature parameter. In all cases, we highlight that reducing the temperature value (x-axes) reduces the average standard deviation (y-axes) across target groups (rows), regardless of how coarse or how granular our aggregation of the personas is (columns).

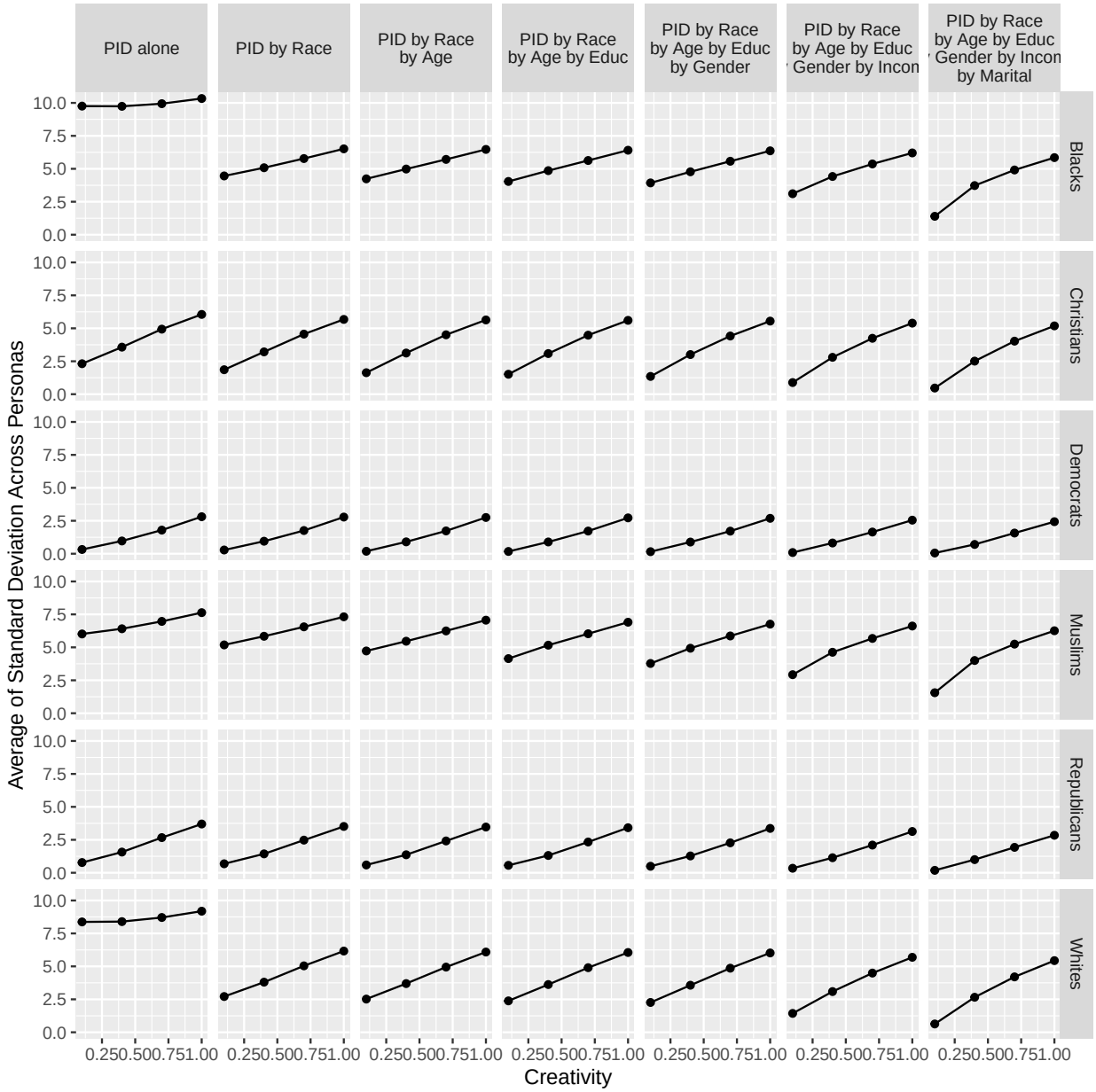

Figure 1: Relationship between temperature hyperparameter (x-axes) and average standard deviation of attitudes (y-axes) by target group (rows) and level of persona aggregation (columns).

3 Replicating Results with Larger Samples

For clarity of exposition, our main results are based on 3,389 ANES respondents who responded to all of the feeling thermometer questions used in the main analyses. However, this restriction threw away many respondents who answered some of the thermometer questions, but not all. If we proceed target group-by-target group, our sample sizes increase to between 7,602 respondents who answered both thermometer questions about Christians and Muslims, to 9,536 respondents who answered both thermometer questions about the Democratic and Republican parties. In the following, we replicate the figures from the manuscript using these larger samples, confirming that – if anything – our substantive conclusions are strengthened with these data. In these data, LLM estimates are an average of 6.5 times as extreme as those generated by humans, and are

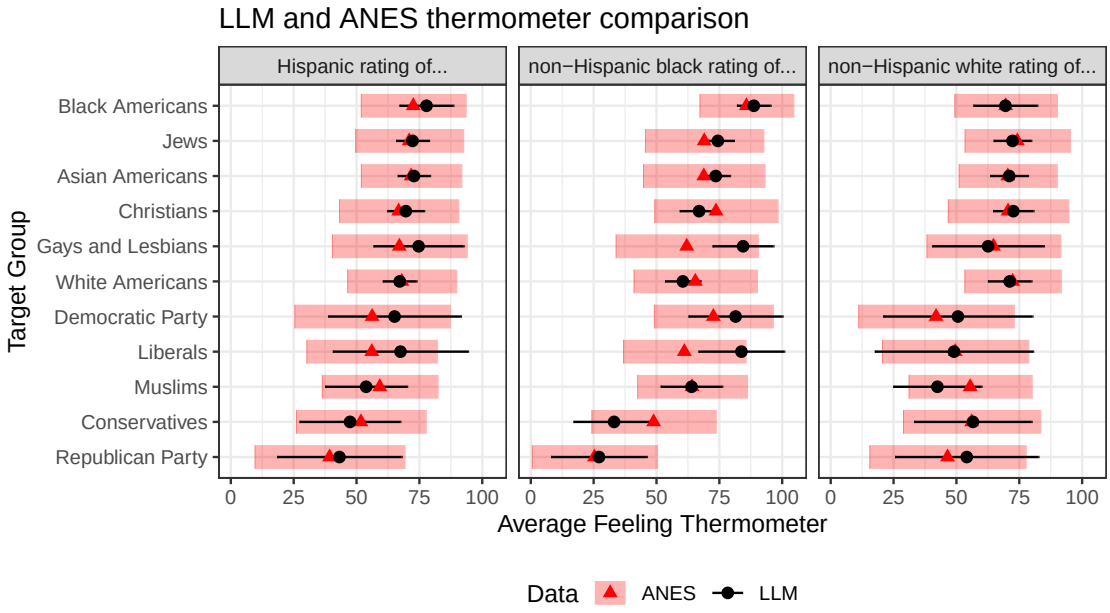

Figure 2: *Replication of Figure 1:* Average feeling thermometer results (x-axis) for different target groups (y-axis) by race of respondent (columns). Average ANES estimates from the 2016 and 2020 waves indicated with red triangles and one standard deviation indicated with thick red bars. LLM-derived averages indicated with black circles and thin black bars. Sample sizes for each group-wise comparison are identical.

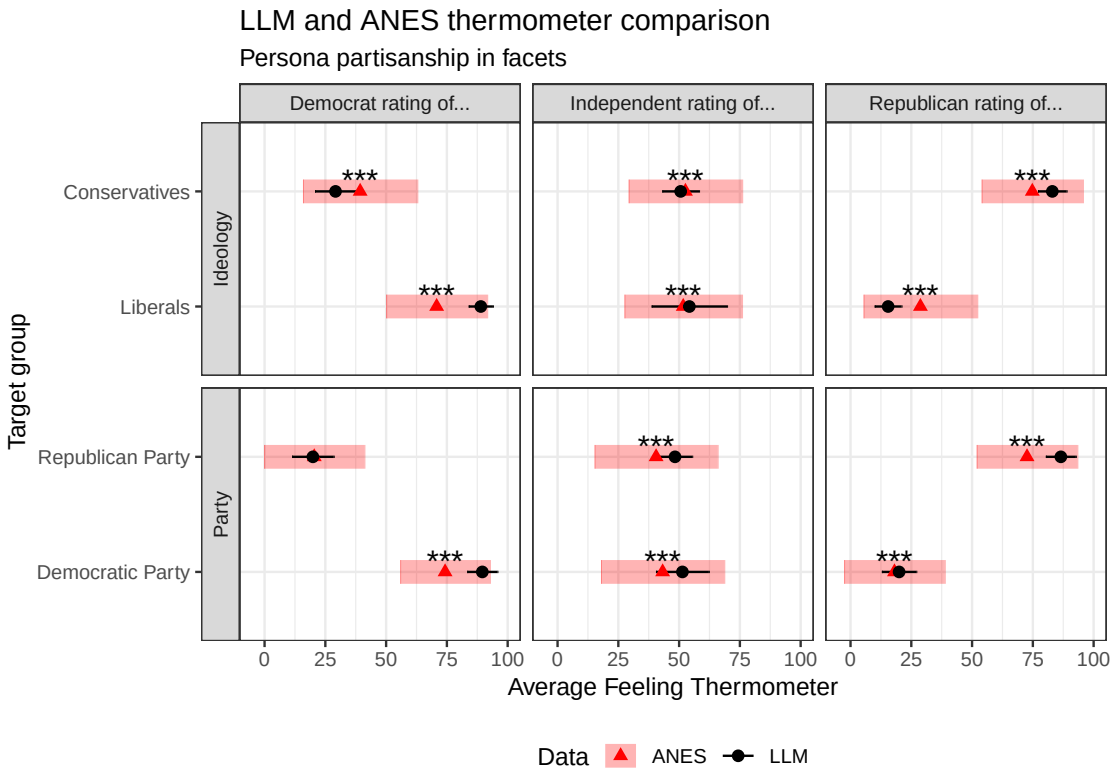

Figure 3: *Replication of Figure 2:* Average feeling thermometer results (x-axis) for different target groups (facets) by party ID of respondent (y-axis). Average ANES estimates from the 2016 and 2020 waves indicated with red triangles and one standard deviation indicated with thick red bars. LLM-derived averages indicated with black circles and thin black bars. Sample sizes for each group-wise comparison are identical. Statistically significant differences indicated with *** = $p < .001$; ** = $p < .01$; * = $p < .05$.

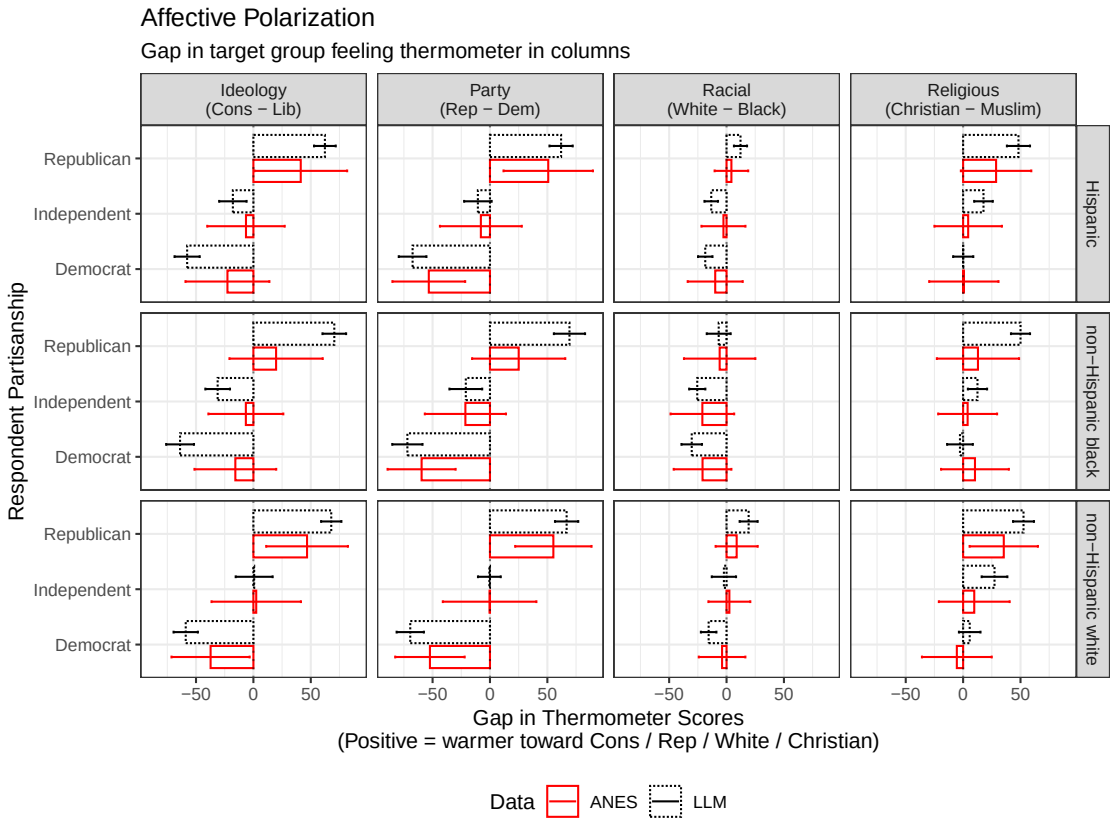

Figure 4: *Replication of Figure 3:* X-axes measure the difference in feeling thermometer ratings between two target groups (columns), by the party ID of the respondent (y-axes) and their race (rows). Black dotted lines indicate results generated by synthetic respondents from the LLM. Red solid lines indicate results generated by real humans from the ANES. Horizontal lines indicate one standard deviation.

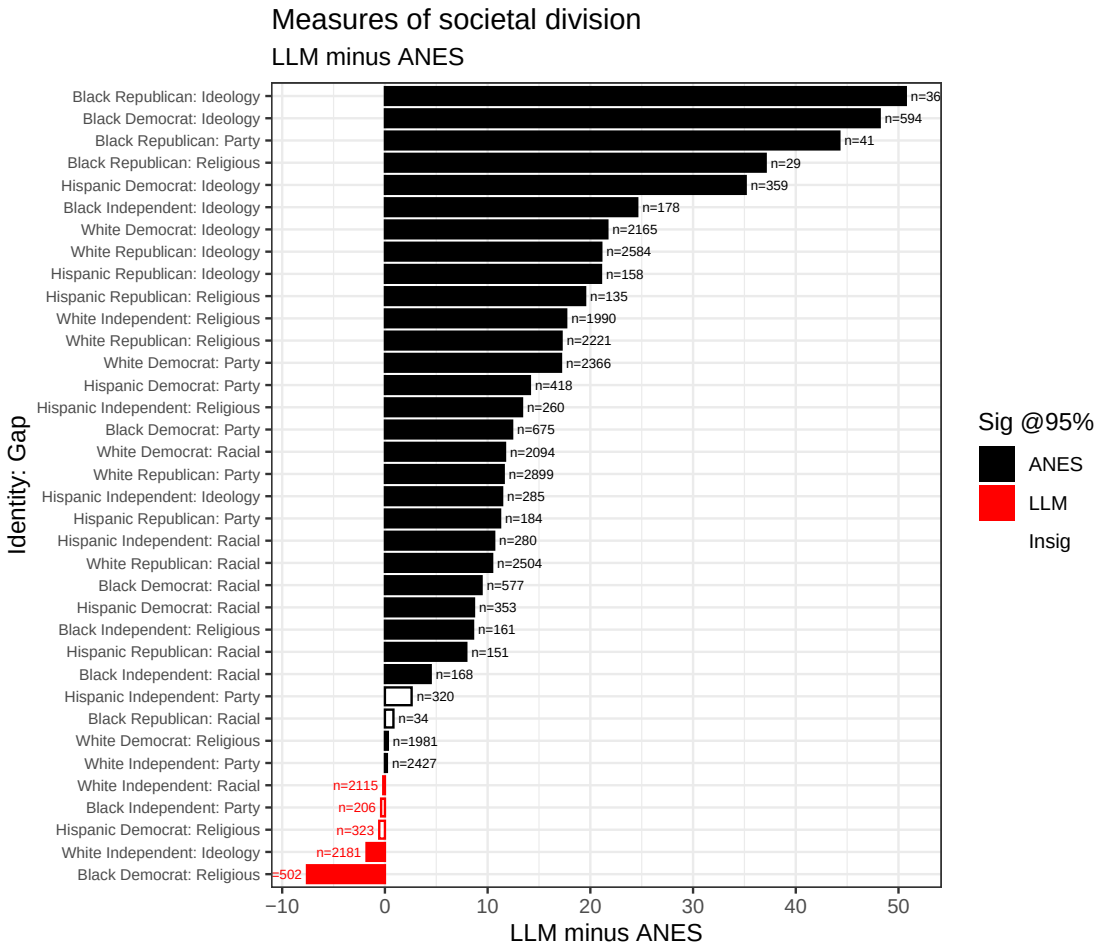

Figure 5: *Replication of Figure 4:* Difference in estimated societal polarization along the dimensions of ideology, partisanship, race, and religion between thermometer gaps estimated from ChatGPT (LLM in black) and from human survey respondents (ANES in red), by race and partisanship of the respondents (y-axis). Solid bars indicate differences between the two data sources that are significant at the 95% level of confidence, while hollow bars indicated statistically insignificant results at this threshold. Number of respondents in each category with ANES responses given by numbers.

4 "Generic" Americans

Our main results are based on prompts to ChatGPT to adopt a specific persona, defined along the characteristics of race, age, gender, education, income, and party ID. Our subsequent analyses then aggregated over different depths of these dimensions to characterize the bias among racial and partisan groups. Here, we instead ask ChatGPT to adopt the persona of the average American and provide the same estimates. Specifically, we instruct ChatGPT to adopt the following generalized identities:

- Basic: person, registered voter
- Party: Democrat, Republican, Independent voter
- Age: 20, 35, 50, 65 year old
- Race: non-Hispanic white, non-Hispanic black, Hispanic
- Gender: male, female
- Income: person making $30,000, $50,000, $80,000, $100,000, $150,000 per year
- Ideology: liberal, moderate, conservative

Each of these identities is not overlaid with other dimensions, meaning we only ask ChatGPT to pretend to be a person living in the USA, a 20 year old living in the USA, a Republican living in the USA, etc. We then characterize how similar these generic identities are to 1) ANES profiles and 2) the richer ChatGPT profiles used in the main analysis. In conducting this analysis, we no longer apply an exact match, and instead simply compare averages between these different sources.

Our first set of results, visualized in Figure 6, compare how similar the ChatGPT generic identities are to each other. Specifically, we calculate the average attitudes toward different outgroups with the "person living in the USA" prompt to the same generated by the "Democrat / Republican living in the USA" prompt, and then subtract the absolute value of the Democrat difference from the absolute value of the Republican difference. As illustrated, the generic American is more similar to the generic Democrat than to the generic Republican across almost all outcomes, and significantly more for groups more associated with Democrats.

To calculate the difference between generic synthetic respondents and real ANES humans, we compare the average feeling thermometer for each synthetic persona to the average feeling thermometer among actual Democrats and Republicans in the ANES. We then plot the difference in these differences in Figure 7, where negative values (meaning that the difference between the LLM and the ANES Democrats is smaller than the difference between the LLM and the ANES Republicans) are indicated in blue, and positive values are indicated in red. Substantively, the plot highlights that, across most personae and across most target groups, ChatGPT's attitudes are more similar to Democrats than to Republicans. Importantly, the only personae where this pattern doesn't hold on average are when we prompt it to adopt the identity of a conservative American, a Republican American, or a non-Hispanic white American. Perhaps even more importantly, ChatGPT's attitudes toward all target groups are more similar to actual Democrats' attitudes than Republicans, with the exception of attitudes toward Jews, the Republican Party, and Muslims. The latter, in particular, has more Republican-leaning attitudes among synthetic ChatGPT persona for every identity excepting women, Hispanics, non-Hispanic Blacks, liberals, and Democrats.

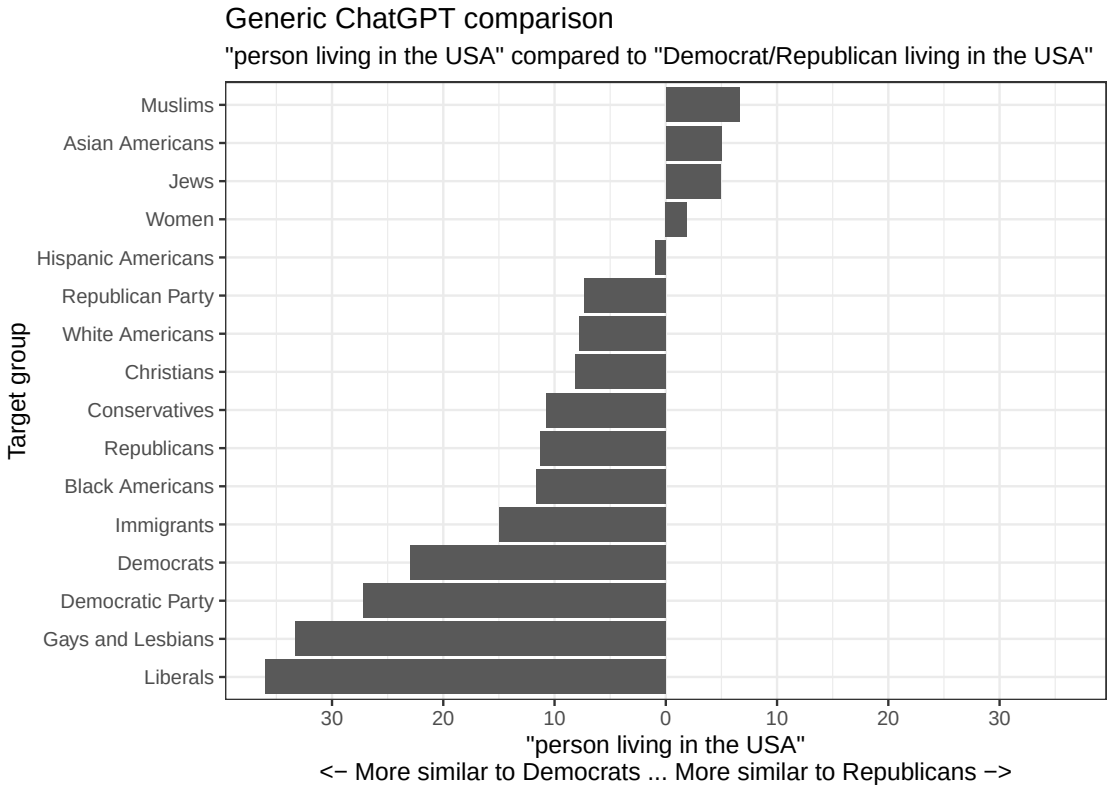

Figure 6: Comparing a generic American ("person living in the USA") with a generic Democrat / Republican, all estimated using ChatGPT. X-axis indicates the difference in the absolute gap between the generic American and the generic Democrat, and the absolute gap between the generic American and the generic Republican. Negative values indicate that the gap between the generic Democrat and the generic American is smaller than the gap between the generic Republican and the generic American, while positive values indicate the opposite.

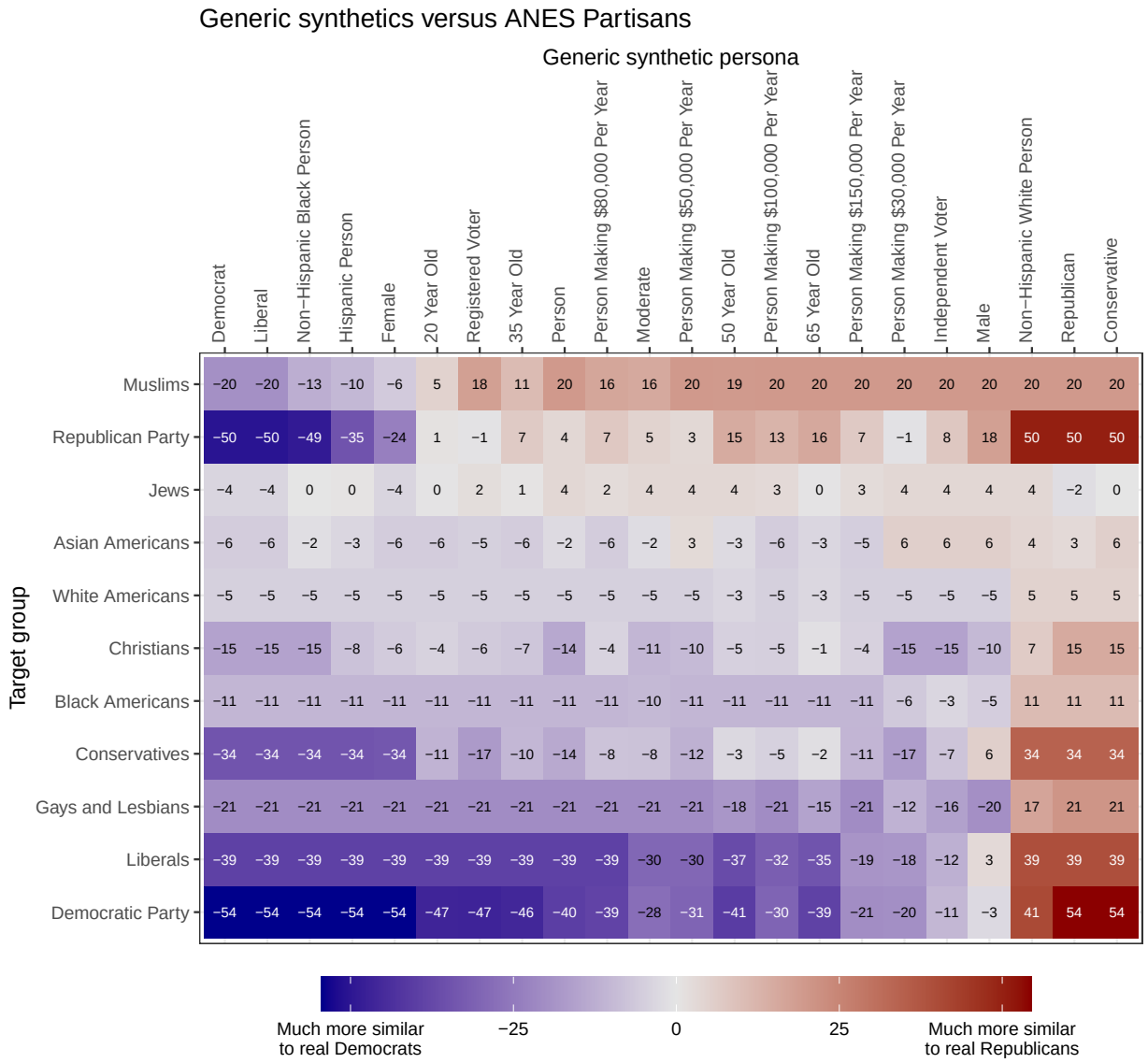

Figure 7: Similarity between generic synthetic ChatGPT respondents (y-axis) and human ANES Democrats (negative values in blue) and Republicans (positive values in red) across a range of target groups (y-axis). Majority of personae are more similar to human Democrats across a majority of target groups.

5 Replicating with different survey

Our main results rely on ANES data, one of the most well-known nationally representative public opinion polls of U.S. politics. We also validate our results using a bespoke survey fielded in 2021 that investigated affective polarization in the context of the Covid-19 pandemic. While the number of target groups is reduced in this survey, we nevertheless are able to implement the same methods described in our manuscript, validating the generalizability of our results to a different period using a different sample from a different polling source.

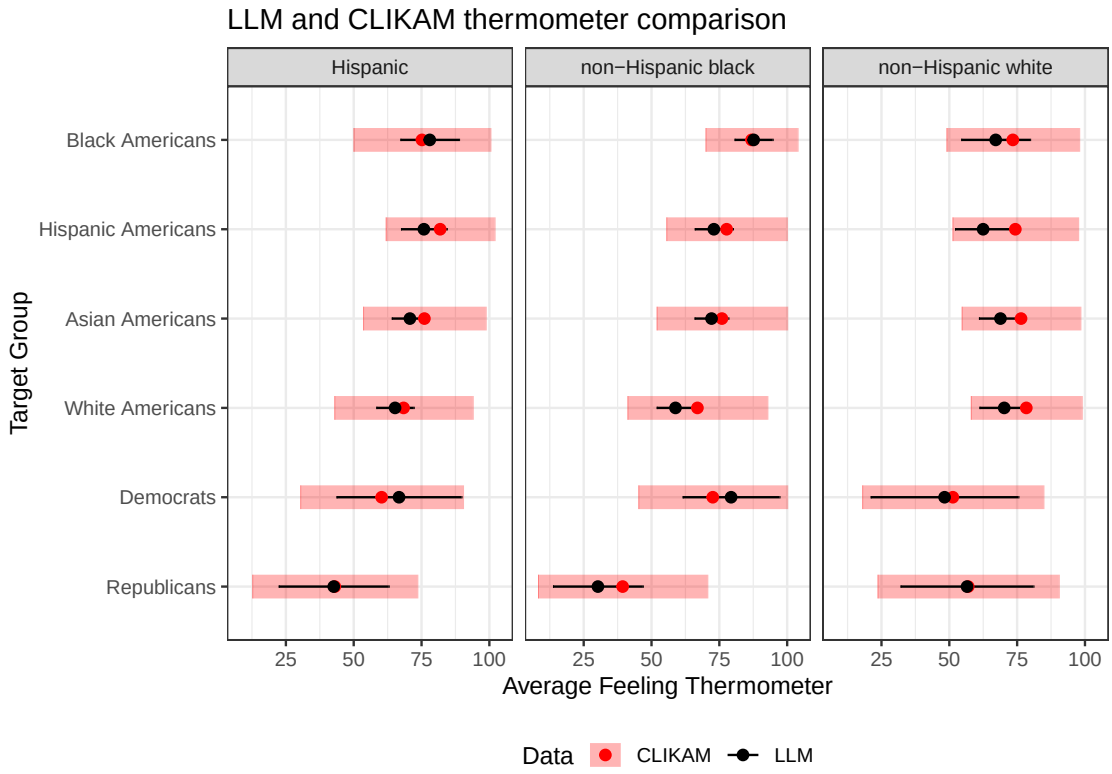

Figure 8: *Replication of Figure 1 in different data:* Average feeling thermometer results (x-axis) for different target groups (y-axis) by race of respondent (columns). Average human estimates from bespoke 2021 survey on affective polarization during Covid-19 indicated with red triangles and one standard deviation indicated with thick red bars. LLM-derived averages indicated with black circles and thin black bars. Sample sizes for each group-wise comparison are identical.

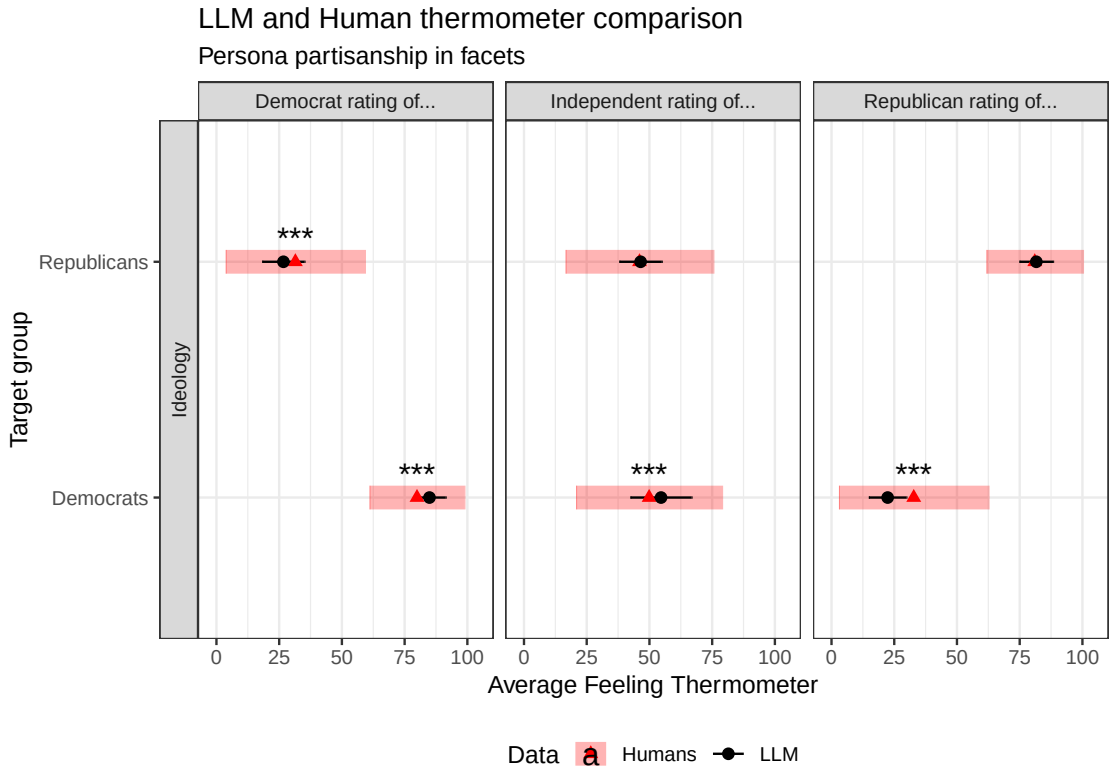

Figure 9: *Replication of Figure 2 in different data:* Average feeling thermometer results (x-axis) for different target groups (facets) by party ID of respondent (y-axis). Average human estimates from bespoke 2021 survey on affective polarization during Covid-19 indicated with red triangles and one standard deviation indicated with thick red bars. LLM-derived averages indicated with black circles and thin black bars. Sample sizes for each group-wise comparison are identical. Statistically significant differences indicated with *** = $p < .001$; ** = $p < .01$; * = $p < .05$.

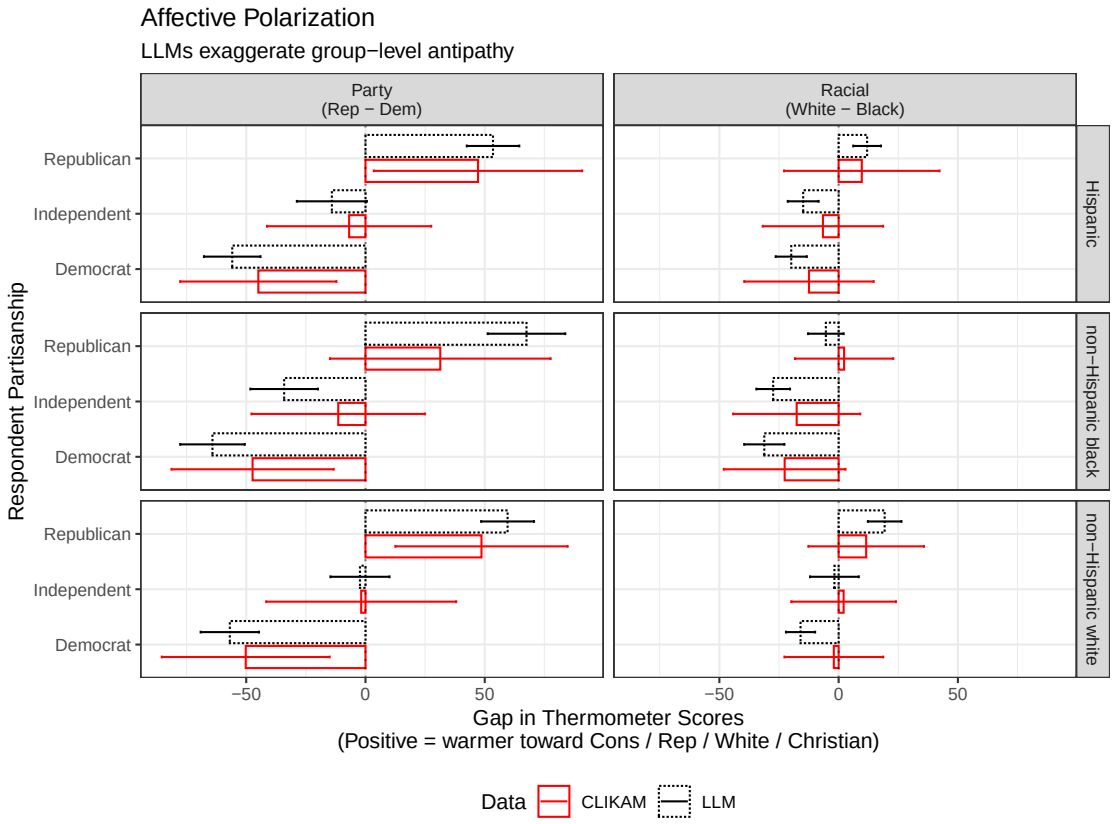

Figure 10: *Replication of Figure 3 in different data:* X-axes measure the difference in feeling thermometer ratings between two target groups (columns), by the party ID of the respondent (y-axes) and their race (rows). Black dotted lines indicate results generated by synthetic respondents from the LLM. Red solid lines indicate results generated by real humans from bespoke survey on affective polarization during Covid-19. Horizontal lines indicate one standard deviation.

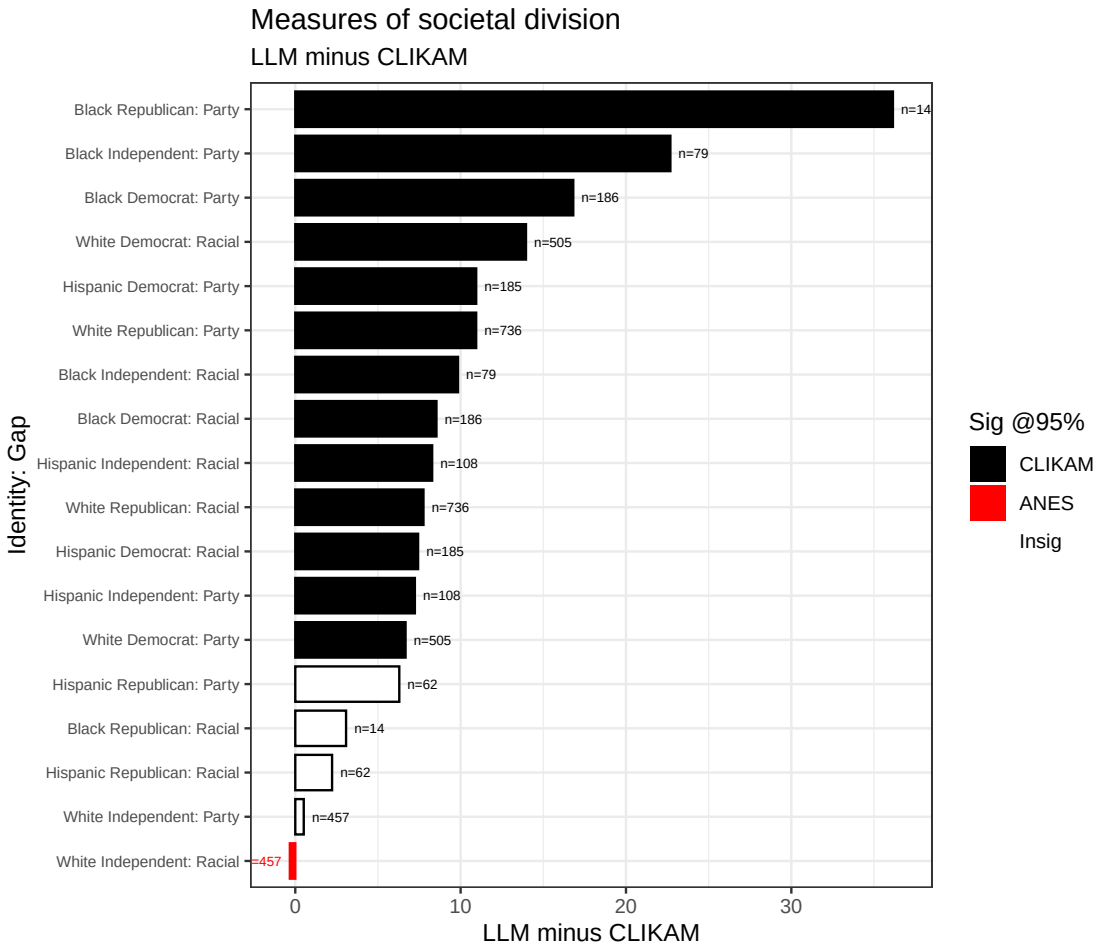

Figure 11: *Replication of Figure 4 in different data:* Difference in estimated societal polarization along the dimensions of ideology, partisanship, race, and religion between thermometer gaps estimated from ChatGPT (LLM in black) and from human survey respondents (bespoke survey in red), by race and partisanship of the respondents (y-axis). Solid bars indicate differences between the two data sources that are significant at the 95% level of confidence, while hollow bars indicated statistically insignificant results at this threshold. Number of respondents in each category with human responses given by numbers.

6 Detailed description of summary statistics

Our main analyses refer to several summary statistics to characterize the degree to which LLM-generated opinions are more extreme (roughly 7 times as much as humans) or more precise (roughly 40% as variable as humans). Here, we break down these numbers for the interested reader.

6.1 Exaggerated Extremism

To calculate the degree to which ChatGPT-generated responses are more extreme than found among human respondents in the ANES, we turn to our matched data where every human respondent in the ANES is matched with a set of random pulls from the ChatGPT API, prompted to adopt the persona of the ANES respondent along the dimensions of age, gender, race, education, income, and partisanship. For each demographic profile defined by these characteristics, we calculate the average feeling thermometer expressions toward the set of target groups asked in the ANES. We then calculate the same average among the synthetic humans generated by ChatGPT which were matched to the real ANES respondents. For each demographic profile, we thus obtain an average feeling thermometer for a given target group estimated by ChatGPT and among the humans sampled by the ANES. Since a thermometer score of 50 captures an indifferent or "neutral" attitude toward a given outgroup, we measure extremism as the absolute difference between the recorded attitude and 50. Our summary statistic of ChatGPT's extremism is thus the ratio of the LLM model's average absolute difference divided by the ANES data's average absolute difference. Ratios greater than 1 indicate that ChatGPT's estimates are more extreme (either more warm or more cool) than real human attitudes, while those less than 1 indicate the opposite. On average, across all covariate profiles and all target groups, synthetic responses gathered from ChatGPT are 4.88 times more extreme than those recorded among real human respondents to the ANES. Broken out by target group, we see consistent evidence that ChatGPT's estimates are always further from 50 on average than the averages found among real humans responding to the ANES.

6.2 Exaggerated Polarization

Turning to the question of antipathy toward outgroups (measured as either affective polarization, partisan sectarianism, or racial or religious antipathy), we pursue a similar exercise. Specifically, we calculate the measure of "polarization" along each dimension, subtracting sentiments towards liberals, Democrats, Blacks, or Muslims from the sentiments towards conservatives, Republicans, Whites, and Christians, respectively. The resulting measures are positive when the respondent is more warm toward stereotypically conservative / Republican groups, and negative when the respondent is more warm toward stereotypically liberal / Democratic groups. Given the matched nature of the data, we are able to calculate these measures using both the real human's answers to ANES survey questions as well as what ChatGPT thinks someone fitting their Demographic profile would say. We then calculate the ratio of the absolute value of the LLM's measure of polarization relative to that found in the ANES for each respondent (again relying on the absolute value). The resulting overall average ratio across all respondents and all groups is just under 7, suggesting that ChatGPT estimates are biased toward out-group antipathy.

However, ratios are sensitive to outliers, particularly in the denominator. If the ANES difference happens to be zero (a likely event given the pressures of social desireability bias in which respondents want to be seen as egalitarian), whatever small difference found in the LLM results will be exaggerated. We therefore plot the raw distributions of polarization across the four measures of partisanship, ideology, race, and religion (y-axis) in Figure 13. As illustrated, ChatGPT's estimates are consistently more polarized than those found among human respondents to the ANES,

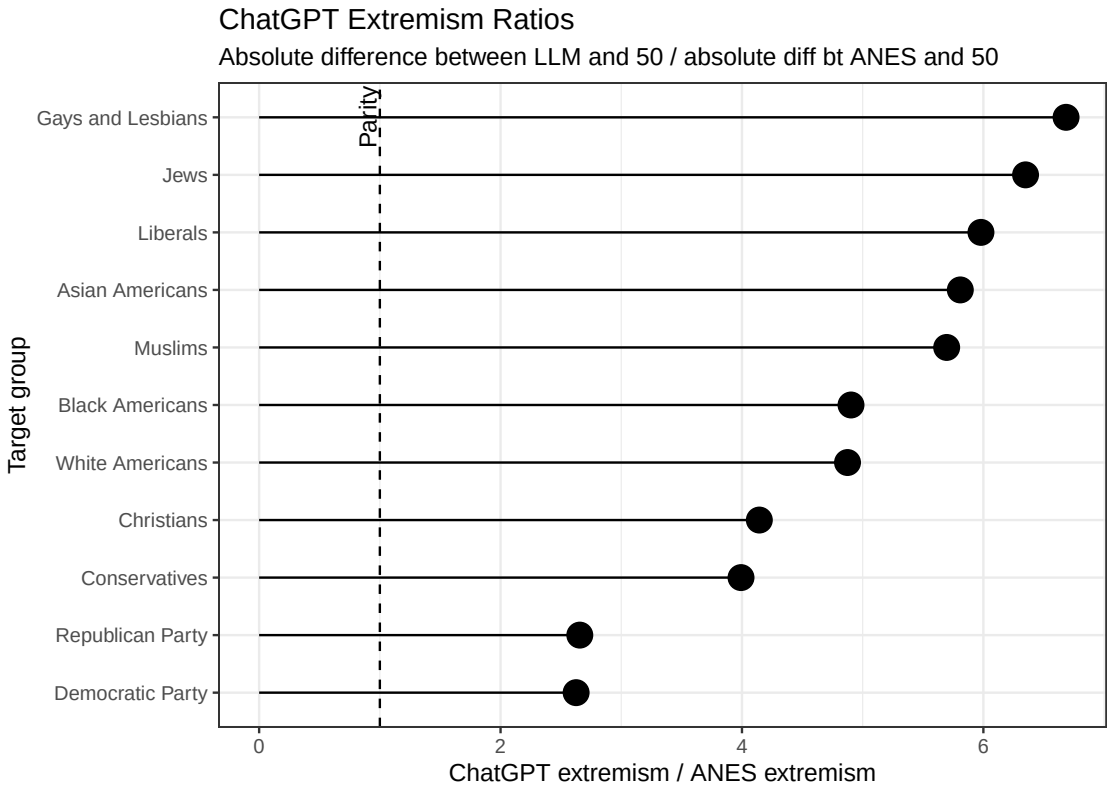

Figure 12: Ratio of ChatGPT extremism (absolute difference between estimate and 50) to ANES extremism. Dashed vertical line indicates parity.

although we note that there are many examples where an ANES respondent's attitudes are more polarized than its synthetic counterparts.

6.3 Exaggerated Confidence

To evaluate the degree to which ChatGPT estimates are less variable (i.e., more confident) than estimates generated by real humans, we start by describing the calculation used in the manuscript. Specifically, we calculate the standard deviation for each race-by-party-by-target group profile in the LLM and ANES data, then divided the latter by the former. We plot the results of this exercise in Figure 14, averaging over race and party ID to calculate a summary measure of how much more confident ChatGPT is in its estimates than the ANES. As above, the vertical dashed line at 1 indicates parity. As illustrated, ChatGPT's standard deviations are consistently between one-quarter and one-third the size of those found among human respondents to the ANES, with an overall average of roughly 0.314.

An alternative way of characterizing ChatGPT's exaggerated precision is by looking across all demographic profiles with two or more respondents in the ANES data, and dividing the ChatGPT standard deviation by that in the ANES. We plot each profile's ratio in Figure 15, sizing the points by the number of observations in each profile, and labeling each outcome by the proportion of profiles whose ChatGPT-derived estimate is more precise than the ANES-derived estimate. Overall, 95% of profiles with two or more ANES respondents that expressed a feeling thermometer toward an outgroup had smaller standard deviations when estimated using ChatGPT relative to

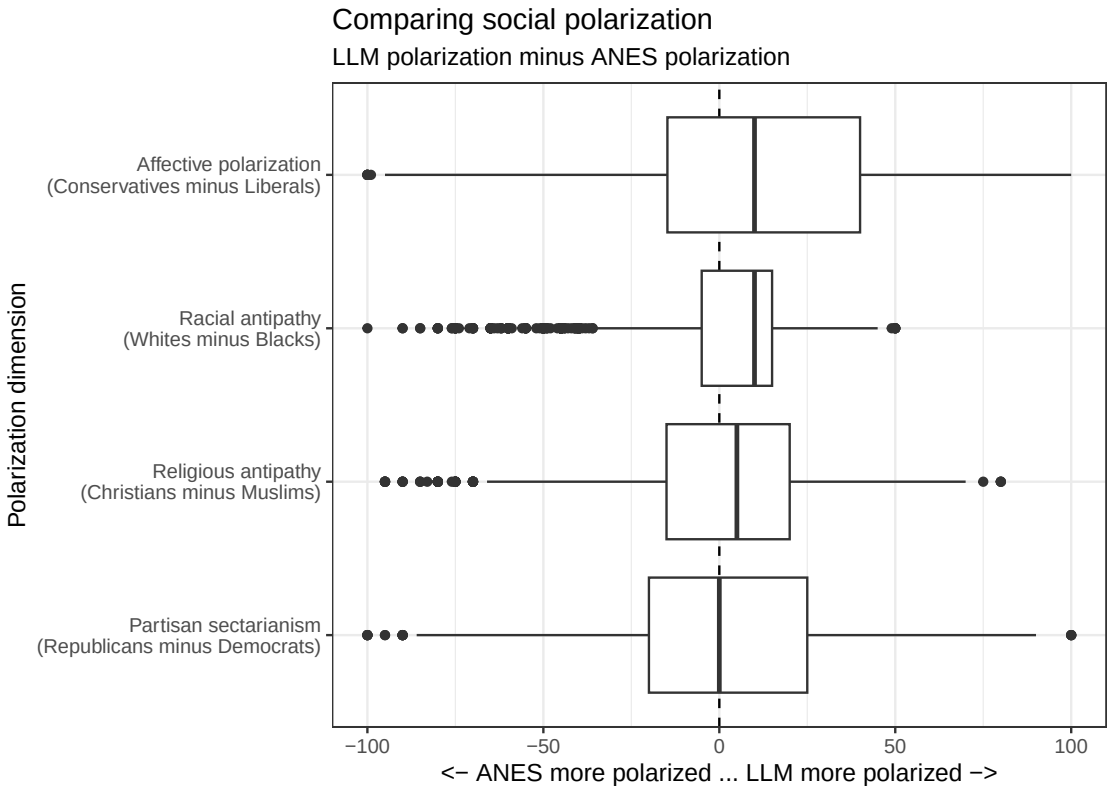

Figure 13: Difference between ChatGPT polarization and ANES polarization.

human respondents to the ANES. And as Figure 15 makes clear, this conclusion is far stronger if we focus on demographic profiles with more ANES respondents.

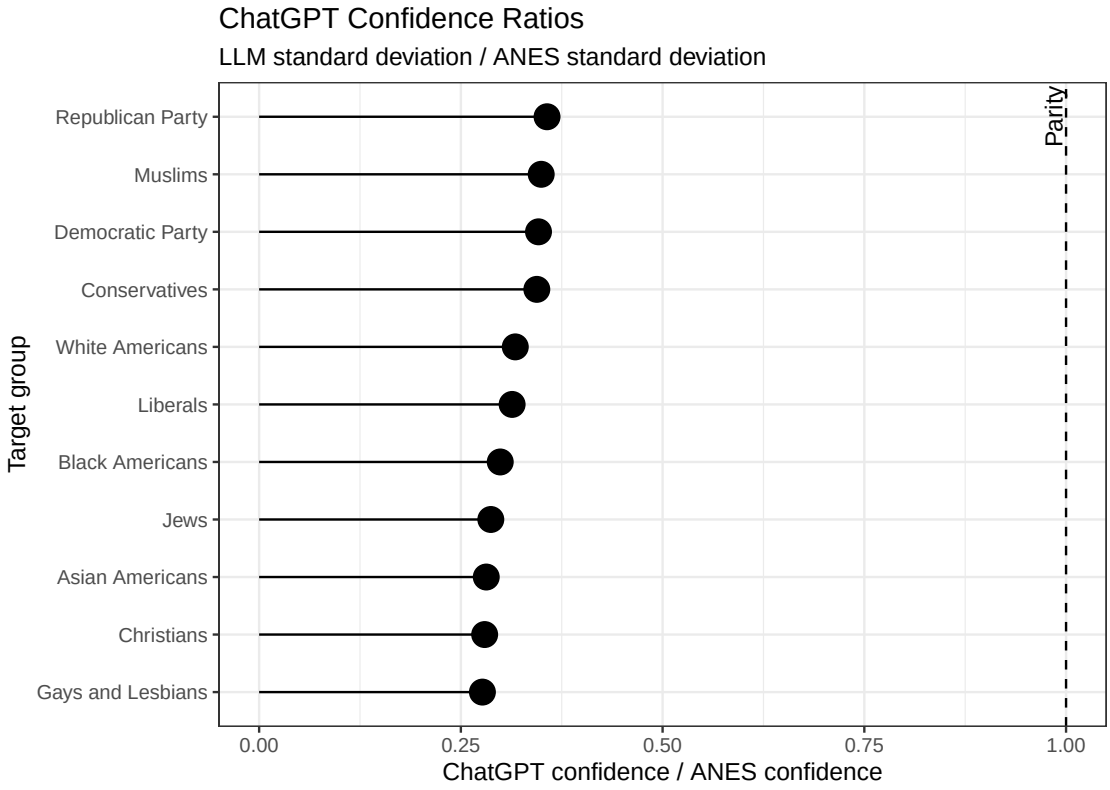

Figure 14: Ratio of ChatGPT confidence (standard deviation across race and party) to ANES confidence. Dashed vertical line indicates parity.

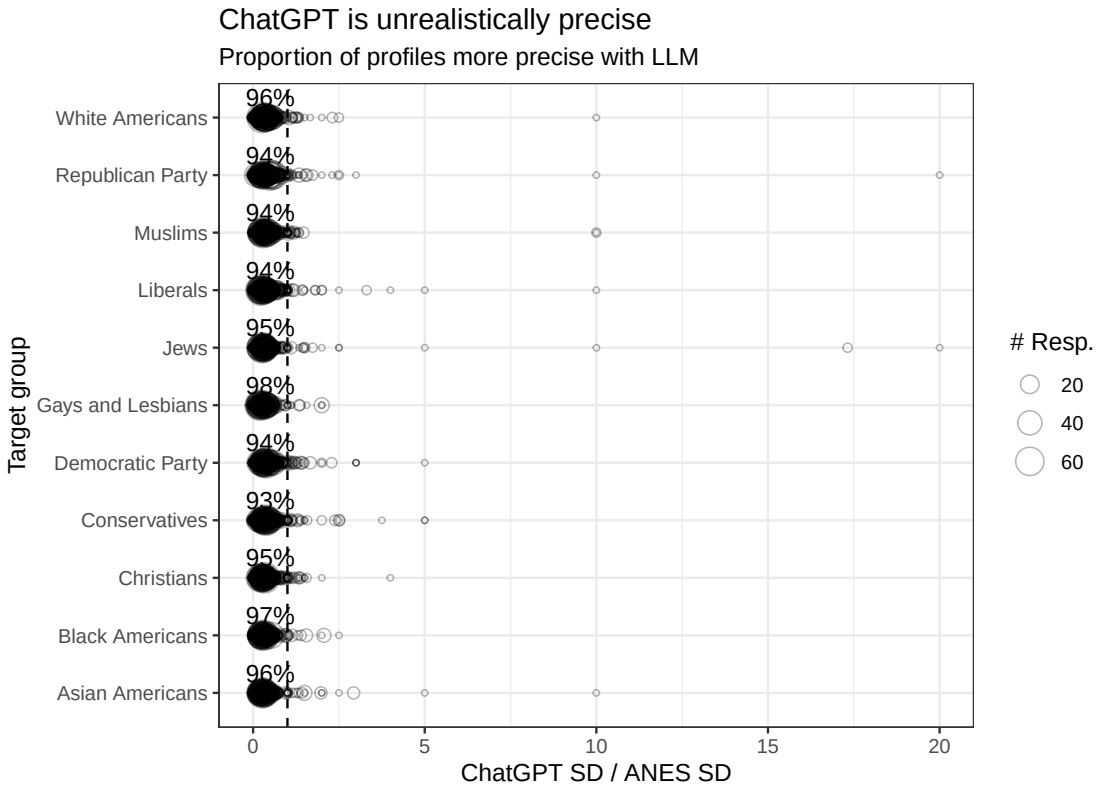

Figure 15: Ratio of LLM standard deviation to ANES standard deviation (x-axis) by demographic profile (circles) sized by total number of respondents. Ratios less than 1 indicate profiles in which the LLM-derived estimate was more precise than the same measure calculated based on the ANES. Text indicates the proportion of profiles per target group (y-axis) that were more precise when estimated via ChatGPT than via ANES.